Máster en Investigación en Informática, Facultad de Informática,
Universidad Complutense de Madrid

Técnicas de visión estereoscópica para determinar la estructura tridimensional de la escena

Proyecto Fin de Máster en Ingeniería Informática para la Industria

Autor: Martín Montalvo Martínez
Director: Gonzalo Pajares Martinsanz
Curso académico 2009/2010

AUTORIZACIÓN DE USO

El abajo firmante, matriculado en el Máster en Investigación en Informática de la Facultad de Informática, autoriza a la Universidad Complutense de Madrid (UCM) a difundir y utilizar con fines académicos, no comerciales y mencionando expresamente a su autor el presente Trabajo de Fin de Máster: “Técnicas de visión estereoscópica para determinar la estructura tridimensional de la escena”, realizado durante el curso académico 2009-2010 bajo la dirección de Gonzalo Pajares Martinsanz en el Departamento de Arquitectura de Computadores y Automática, y a la Biblioteca de la UCM a depositarlo en el Archivo Institucional E-Prints Complutense con el objetivo de incrementar la difusión, uso e impacto del trabajo en Internet y garantizar su preservación y acceso a largo plazo.

Firma del autor:

Martín Montalvo Martínez

RESUMEN

En este trabajo se realiza un estudio sobre la efectividad de una serie de métodos de correspondencia estereoscópica. La correspondencia estereoscópica constituye uno de los pasos esenciales dentro de la visión estereoscópica en los sistemas robotizados, de ahí su importancia. El objetivo se centra en el estudio de la viabilidad de los mismos de cara a su implementación en sistemas estereoscópicos que han de operar en entornos de exterior y bajo condiciones del entorno adversas. La motivación del trabajo proviene de la necesidad derivada de una serie de proyectos de investigación dentro de las actividades del grupo ISCAR. En este trabajo se han realizado diversas pruebas experimentales orientadas a la identificación de los métodos más prometedores en el ámbito de la correspondencia estereoscópica con la finalidad indicada. Se han estudiado varias técnicas existentes en la literatura y se han establecido las pautas a seguir en el futuro a tenor de los resultados obtenidos para su implementación en sistemas reales.

PALABRAS CLAVE

Visión estéreo, mapa de disparidad, correspondencia estéreo, mapas cognitivos borrosos, oclusión, segmentación de imágenes

ABSTRACT

In this work we have studied several stereovision matching approaches with the aim of testing its effectiveness. The main step in robotized systems, equipped with stereovision, is the correspondence, here is its relevance. The goal of this work is focused on the study of the viability of such methods with the aim that they can be implemented in stereoscopic vision-based systems working in adverse outdoor environmental conditions. This work is motivated because the ISCAR group is currently working in several research projects where the stereovision is a crucial system. In this work several experimental tests have been carried out oriented toward the identification of the most promising correspondence methods with the above expressed goal. Several existing approaches in the literature have been studied and, as a result, some guidelines have been established based on the results reported, so that the research is oriented toward future implementations in real systems.

KEYWORDS

stereovision, disparity map, stereo matching, fuzzy cognitive maps, occlusion, image segmentation

ÍNDICE

CAPÍTULO 1

1. Introducción

1.1. Identificación del problema

Hoy en día, tanto las cámaras como los computadores están totalmente implantados en la sociedad, y su uso es, en muchas ocasiones, indispensable. Las cámaras se utilizan en diversos ambientes y para distintas finalidades, destacando por ejemplo: seguridad (vigilancia de edificios, museos), comunicaciones (videoconferencias), entretenimiento (videoconsolas, cámaras fotográficas, cámaras de video), etc. Y si las cámaras se utilizan de forma generalizada, el uso de los computadores supera al de éstas en gran medida. Debido al desarrollo que ha experimentado la industria electrónica, se dispone de microprocesadores muy pequeños y baratos, lo que ha conducido a que actualmente se encuentren sistemas empotrados en casi todos los elementos de uso cotidiano: teléfonos móviles, coches, videoconsolas, electrodomésticos, monitores, cajeros automáticos, decodificadores para la recepción de televisión, etc. Algunos de estos dispositivos contienen más de un microprocesador en su interior.

Un campo en el que confluyen cámaras y ordenadores es, como su nombre indica, el de la Visión por Computador y dentro de éste la visión estereoscópica. Como fácilmente se puede deducir, las imágenes son bidimensionales mientras que la escena cotidiana es tridimensional. Esto significa que entre el paso de la escena, que es la realidad, a la imagen se ha perdido lo que denominamos la tercera dimensión. La visión estereoscópica constituye un procedimiento más para la obtención de esa tercera dimensión perdida y a partir de ella en la medida de lo posible la obtención de la forma de los objetos en la escena. Nuestro sistema visual humano es capaz de

percibir en tres dimensiones y además es estereoscópico, constituido por dos ojos, ello es lo que ha hecho que los sistemas estereoscópicos artificiales utilicen al menos, dos imágenes distintas de la misma escena. Con ellas se puede llegar a determinar la distancia a la que se encuentra un objeto cualquiera, contenido en las dos imágenes, respecto del observador. Las cámaras se utilizan para captar las imágenes y el computador se requiere para realizar los cálculos que determinan la distancia al observador. El sistema visual humano percibe la escena y el cerebro procesa la información, ésta es la analogía entre el sistema artificial y el biológico. Eso sí, el artificial podría decirse que se encuentra a años luz del humano y en absoluto consigue los resultados del primero. Centrándonos ya sobre el sistema artificial, la información sobre las distancias a las que se encuentran los objetos situados en la escena se generan en forma de una estructura conocida técnicamente como mapa de disparidad, que no es ni más ni menos que una representación de las diferentes profundidades a las que se encuentran los objetos respecto de la ubicación de las cámaras. Posteriormente, mediante una simple relación geométrica por semejanza de triángulos y conocidos algunos parámetros de las cámaras tales como la separación existente entre ellas y las distancias focales de sus sistemas ópticos es posible determinar las distancias buscadas. La obtención del mapa de disparidad requiere realmente la identificación del mismo punto u objeto en las dos imágenes y que representa la misma entidad física en la escena tridimensional. Pues bien, al proceso por el cual se llega a identificar en sendas imágenes esa misma entidad tridimensional se le conoce como correspondencia estereoscópica. En todo proceso de visión estereoscópica la correspondencia constituye el verdadero problema al que se le ha dedicado un altísimo porcentaje de investigación en el campo de la visión estereoscópica, encontrándose en el momento actual totalmente abierto a la investigación. Esta es la razón por la que este trabajo se centra en la estudio de métodos y procedimientos encaminados a tratar de dar una solución al mismo.

1.2. Motivación y objetivos

1.2.1. Motivación

El trabajo que se presenta tiene su origen en las actividades de investigación planteadas dentro del grupo ISCAR (2010) destacando los siguientes proyectos actualmente en vigor:

1) Plataforma de Planificación, Simulación y Sistema de Vigilancia, Búsqueda y Rescate en el Mar mediante colaboración de Vehículos Autónomos Marinos y Aéreos (DPI2009-14552-C02-01), 2009-2012, perteneciente al Plan Nacional de I+D+i.

2) FONCYCIT (Unión Europea-México) (2009-2011) Análisis de Imágenes para el Control de Robots Autónomos, con participación de la Universidad Complutense, el Instituto Politécnico Nacional de Méjico, la Universidad de Guadalajara en México y la Universidad Libre de Berlín.

Con anterioridad, el grupo ha desarrollado actividades de investigación en el marco de otros dos proyectos en el ámbito aeroespacial con la empresa TCP Sistemas e Ingeniería titulados:

1) AUTO-ROVER: estudio de Autonomía basada en imágenes para Rover de exploración planetaria.

2) Visión Estereoscópica para AUTOROVER: Investigación aplicada de Autonomía basada en imágenes para “Rover” de Exploración planetaria.

En todos ellos la visión estereoscópica constituye uno de los componentes esenciales ya que su orientación es hacia la navegación autónoma de vehículos, tanto aéreos como terrestres no tripulados, los cuales están dotados con los correspondientes sensores ópticos de visión artificial. En todos ellos el principal problema que se plantea es el de la correspondencia estereoscópica, máxime teniendo en cuenta que todos ellos operan en entornos de exterior donde los

problemas relacionados con el procesamiento de las imágenes se complican enormemente en relación a los entornos de interior.

Como ya se ha mencionado previamente, el proyecto que se presenta se orienta hacia la identificación de los problemas relativos a la correspondencia estereoscópica, donde la identificación de un algoritmo apropiado, junto con toda la problemática de los entornos de exterior, constituyen los fundamentos del proyecto.

1.2.2. Objetivos

En base a lo anteriormente expuesto, los siguientes objetivos son los que se plantean dentro del proyecto de investigación que se presenta:

1) Aprender a manejar referencias bibliográficas, así como la forma de abordar las investigaciones.
2) Identificar métodos de correspondencia estereoscópica existentes en la literatura para la obtención de la estructura tridimensional de la escena.
3) Determinar los métodos más apropiados para el cómputo de disparidades en los entornos de exterior donde va a operar el sistema.
4) Analizar las ventajas e inconvenientes de cada uno de ellos, así como su problemática.
5) Determinar los métodos para la mejora del mapa de disparidades obtenido en primera aproximación.
6) Realizar un aporte de carácter investigador mediante las conclusiones finales.

Como aportación a la investigación, en el presente trabajo se realiza un estudio comparativo sobre el comportamiento de los métodos de correspondencia estereoscópica más prometedores desde el punto de vista de las aplicaciones a las que se destinan. Dicho estudio se realiza sobre imágenes con disparidades conocidas, con el fin de poder establecer las bases de aplicación a entornos reales con la complejidad que les caracteriza.

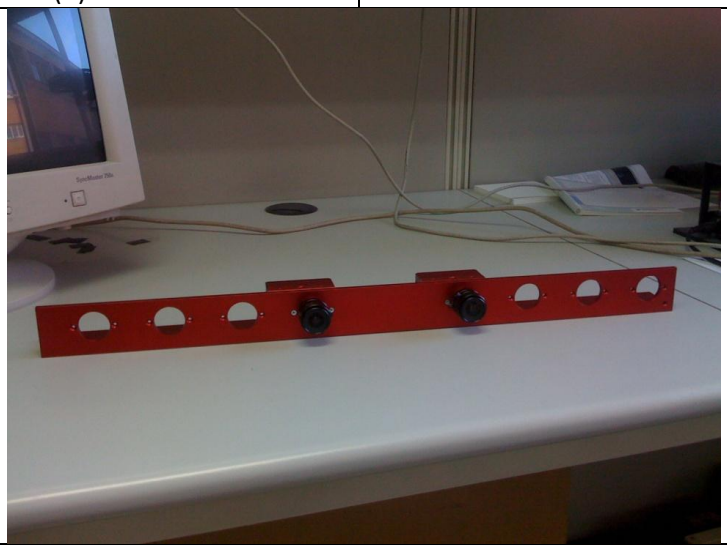
Para dar idea de las posibilidades de aplicación de las técnicas que constituyen el objeto de la investigación, en la figura 1.1 se muestran tres sistemas estereoscópicos reales sobre los que se pretende instalar los procedimientos de correspondencia estereoscópica aquí estudiados. Estos sistemas se encuentran actualmente operativos en los laboratorios del grupo de investigación ISCAR (2010) sobre los que está prevista la continuación de los trabajos de investigación en el futuro. En esta línea se pretende revisar conjuntamente los resultados aquí obtenidos con los procedentes de otros investigadores en el ámbito de investigación del grupo bajo la cobertura de los proyectos mencionados en la sección previa. El objetivo consiste en aunar los esfuerzos y coordinar las acciones hacia la consecución de las tareas planteadas en dichos proyectos.

Continuando con los sistemas estereoscópicos anteriormente mencionadas, en la Figura 1-1(a) se muestra el robot Surveyor, que se trata de un producto comercial equipado con un sistema *wifi* que permite transmitir las imágenes al computador y desde éste recibir órdenes de desplazamiento autónomo. En las Figura 1-1(b) y (c) se muestra otro sistema estereoscópico montado sobre un trípode y construido íntegramente en el Laboratorio del Grupo ISCAR dotado con sendas cámaras SCA 140017FC de Basler en color y resoluciones de 1390x1038, fijadas sobre un soporte horizontal. La conexión al computador se realiza también mediante Fire Wire IEEE 1394. Finalmente, en la Figura 1-1(d) se muestra el tercero de los sistemas, suministrado por la empresa VIDERE (2010), y dotado con dos cámaras situadas sobre un soporte rígido con diversos orificios con capacidad para variar la posición relativa de las cámaras en el rango de 18 a 60cm. Este sistema se conecta al computador también a través del estándar Fire Wire IEEE 1394, proporcionando imágenes en color con resoluciones de hasta 1024x860.

(a)

(b)

(c)



(d)

Figura 1-1: Distintos sistemas de visión estereoscópica; (a) Robot "Surveyor" dotado con un sistema de visión estereoscópica; (b) y (c) vistas del sistema estereoscópico montado en el laboratorio del grupo ISCAR; (d) sistema de base variable de VIDERE

1.3. Metodología

El esquema general propuesto para la consecución de los objetivos sigue básicamente los siguientes pasos:

1) En primer lugar se realiza una revisión bibliográfica sobre los métodos descritos en la literatura.
2) Se seleccionan los métodos más apropiados para nuestra problemática.
3) Se estudia su comportamiento sobre imágenes sintéticas y reales de las que se posee información cierta sobre los mapas de disparidad proporcionada por los laboratorios que las han generado.
4) Se estudian los resultados obtenidos sobre las mismas.
5) Se elabora la base para su viabilidad en los entornos reales.

1.4.Organización del trabajo

La memoria se organiza en capítulos, de forma que en el capítulo dos se realiza la síntesis relativa al estudio bibliográfico llevado a cabo sobre revisión de metodologías relacionadas con la visión estereoscópica desde la perspectiva de la correspondencia. En el capítulo tres se describen los métodos considerados como apropiados para resolver el problema de la correspondencia. En el capítulo cuatro se analizan los resultados obtenidos con los mismos sobre el conjunto de imágenes disponibles. Finalmente, en el capítulo cinco se exponen las conclusiones y trabajo futuro.

CAPÍTULO 2

2. Estado del arte

2.1.Introducción

En este capítulo se realiza una revisión de los métodos existentes en la literatura, utilizados para establecer las correspondencias como paso previo para determinar la distancia a la que se encuentran los objetos de la escena respecto del sistema de visión estereoscópica artificial y en relación al observador. Con esta revisión se analizan cuáles son las líneas y tendencias que se siguen, para de esta forma determinar cuáles son los métodos más apropiados en relación a los objetivos planteados.

Los métodos de visión estereoscópica caen dentro de los denominados métodos pasivos, este tipo de métodos no interfieren en la escena que se está analizando. En el otro extremo se encuentran los métodos activos en los que se actúa de alguna forma sobre la escena, ya sea mediante su iluminación o enviando un haz energético. Como ejemplo de método activo estaría el telémetro láser, que mediante el envío de un rayo de esta naturaleza hacia un objeto permite determinar su distancia, ya sea por el tiempo empleado por la onda en llegar al objeto y volver a la fuente, midiendo el cambio de fase experimentado, o la variación en la intensidad de la señal que se refleja (Cassinelli, 2005).

Si bien, anteriormente se ha mencionado que los sistemas de visión estereoscópica artificial utilizan dos o más cámaras, en la mayoría de los métodos y procedimientos existentes en la literatura consultada se hace uso de modelos con dos

cámaras. El modelo de dos cámaras toma como referencia el propio modelo biológico de estereovisión, donde gracias a la distancia existente entre los dos ojos se puede establecer la distancia, esto es la tercera dimensión, a la que se encuentran los objetos. El hecho de que los ojos estén separados entre sí una determinada distancia produce que se obtengan imágenes desplazadas de la misma escena en sendos ojos, es decir la imagen de un ojo es prácticamente la misma que la del otro, pero desplazada una distancia inversamente proporcional a la distancia entre los ojos y los objetos. Las instantáneas que perciben los ojos pertenecen a la misma escena, aunque debido a que éstos no se encuentran situados en el mismo emplazamiento, la perspectiva difiere y por ello estas instantáneas son ligeramente distintas. También el campo visual puede llegar a ser algo diferente en cada ojo, si bien el ángulo total de visión resulta ser prácticamente el mismo, el límite izquierdo del ojo izquierdo se amplia, reduciéndose el límite derecho, en comparación con el ojo derecho, mientras que en el ojo derecho ocurre lo contrario, se amplía el límite derecho y se reduce el límite izquierdo.

Desde el punto de vista artificial, para obtener las imágenes desplazadas que permitan reconstruir la escena tridimensional existen dos aproximaciones básicamente:

- Utilizar dos o más cámaras todas ellas alineadas y separadas una cierta distancia. Esta distancia puede ser cualquiera, depende de la aplicación, pero debe ser conocida con exactitud.

- Utilizar únicamente una cámara móvil. Esta cámara debe ser capaz de desplazarse en línea recta y tomar imágenes mientras realiza este desplazamiento. Al igual que en el caso anterior, se debe conocer el desplazamiento de forma precisa que ha realizado la cámara entre la captura de dos imágenes consecutivas.

En este estudio se trabajará con el primer método, por tanto utilizando dos cámaras alineadas y separadas una distancia fija. Un sistema estereoscópico basado en cámara móvil es el propuesto por Moravec (1997), donde un robot móvil porta una cámara con la que se toma una serie de imágenes. Cada vez que se quiera tomar

imágenes, el robot se detiene de forma que la cámara se desplaza de izquierda a derecha tomando nueve imágenes, este desplazamiento se realiza sobre distancias prefijadas y en el sentido del eje horizontal del sistema óptico.

Otro aspecto inherente a los sistemas de visión estereoscópica es su geometría, pudiéndose optar por una geometría con los ejes ópticos paralelos o convergentes. El sistema visual humano trabaja fundamentalmente con ejes convergentes de forma que enfoca los ojos hacia los objetos de interés. Cuando el objeto está próximo se produce la convergencia de ejes sobre dicho objeto, mientras que si el objeto se sitúa en la lejanía prácticamente no existe convergencia de los ojos, pudiéndose en este caso decir que los ejes ópticos se sitúan en paralelo. Podemos utilizar el sistema de visión humano para comprobar algunas características, ya que ha sido ampliamente estudiado, tanto en el campo de la biología como en el campo de la medicina. Así podemos comprobar por nosotros mismos el desplazamiento que sufren dos imágenes de la misma escena al cambiar la perspectiva. Si con un ojo tapado y fijándonos en un objeto, nos destapamos dicho ojo y nos tapamos el que teníamos descubierto, podremos comprobar el desplazamiento relativo que sufre el objeto al ser observado mediante un ojo (se puede asimilar a una cámara) y mediante el otro. Este desplazamiento es menor cuanto más alejado esté el objeto de nosotros y será mayor cuanto más próximo se encuentre. Para poder aprovechar los conocimientos que se poseen sobre los modelos de ejes ópticos convergentes, se han hecho algunos esfuerzos en visión artificial para desarrollar sistemas que los simulen, a semejanza del modelo biológico, como es el caso de Krotkov (1989) o Krotkov *et al.* (1990). En estos modelos se consigue centrar el foco de atención en uno u otro objeto mediante la convergencia y divergencia de los ejes ópticos. Cuando se enfoca una zona de interés, el resto de las zonas resultan borrosas y no se aprecian correctamente. Como se ha mencionado previamente, al centrar la vista en un objeto lejano, los ejes ópticos divergen hasta situarse prácticamente paralelos, y los objetos cercanos se contemplan con dificultad. Y al contrario cuando centramos nuestra atención sobre objetos cercanos, los ejes convergen, en mayor medida cuanto más cercano esté el objeto, pudiendo observar con claridad estos objetos y con dificultad los lejano. Sin embargo,

el modelo de ejes ópticos paralelos es el más ampliamente utilizado en los sistemas de visión artificial, y es sobre el que se basa el desarrollo del presente trabajo.

2.2.Pasos en el proceso de la visión estereoscópica

Como probablemente el lector haya podido intuir a partir de lo expuesto hasta aquí, cuando se trabaja en visión estereoscópica se trabaja con la adquisición de imágenes y con la modelación de las cámaras. Una descomposición en mayor detalle del proceso de visión estereoscópica, entre los que estarían los dos anteriormente mencionados, fue realizada por Barnard y Fischler (1982). Según esta descomposición, el proceso completo de visión estereoscópica contempla seis pasos principales, a saber:

1. Adquisición de imágenes
2. Modelado de la cámara (geometría del sistema)
3. Extracción de las características
4. Correspondencia de las imágenes (características)
5. Determinación de la distancia (profundidad)
6. Interpolación, cuando sea necesaria

Los pasos anteriores son de naturaleza secuencial con el orden en el que aparecen sobre la figura 2.1.

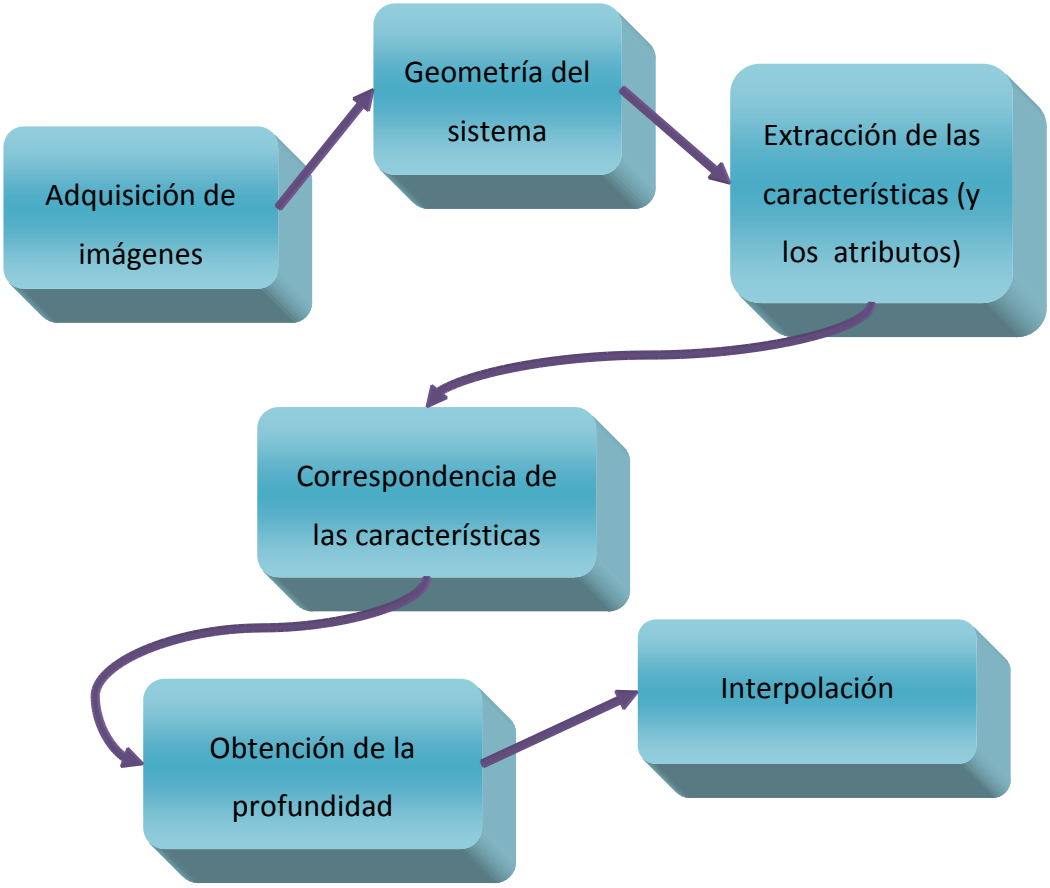


Figura 2-1: Principales pasos en la visión estereoscópica

De estos pasos, al que se reconoce ampliamente como el más complejo de resolver, y que depende clara y fuertemente de la elección de características realizada (extracción de características), es el número cuatro, correspondencia de las imágenes (Cochran y Medioni, 1992).

2.2.1. Adquisición de imágenes

La adquisición de imágenes puede realizarse de muchas formas distintas. Por ejemplo, las imágenes pueden ser tomadas simultáneamente en el tiempo, o mediante intervalos de tiempo con una duración determinada. Pueden ser tomadas desde localizaciones y direcciones ligeramente distintas o radicalmente diferentes. Si las imágenes son adquiridas con una diferencia temporal grande, influye el momento del día en las que fueron capturadas, las condiciones atmosféricas y cualquier elemento que haya cambiado la escena que se está considerando. El factor más determinante en la adquisición de las imágenes es el tipo de aplicación para la cual se quiere obtener la

estructura de la escena tridimensional. No es lo mismo considerar una aplicación basada en un vehículo aéreo tomando fotografías con una única cámara con fines cartográficos, donde las imágenes suelen ser de una baja resolución, apareciendo distintos tipos de terrenos (texturas), que una aplicación para un vehículo terrestre autónomo capaz de evitar obstáculos, donde las imágenes deben ser de una resolución mayor, se deben identificar objetos más que regiones y se suelen tener dos cámaras que toman imágenes simultáneamente. Normalmente las aplicaciones determinan el tipo de escena sobre la que trabajan. Se puede realizar una clasificación de las escenas en dos grupos: escenas con elementos realizados por el hombre, como edificaciones y carreteras; y escenas que contienen únicamente elementos naturales y superficies, como pueden ser montañas, terrenos lisos u ondulados, follaje y agua.

2.2.2. Modelo geométrico de la cámara

Un modelo de cámara es una representación de los atributos geométricos y físicos más importantes de las cámaras utilizadas para la visión estéreo. Este modelo puede tener una componente relativa, la cual relaciona el sistema de coordenadas de una cámara con el de la otra, y es independiente de la escena, y también puede tener una componente absoluta, la cual relaciona el sistema de coordenadas de una de las cámaras con un sistema de coordenadas fijo de la escena. El modelo en el que nos centraremos es uno con dos cámaras que tienen sus ejes ópticos paralelos, siendo la distancia que los separa la *línea base*. Quedando sus ejes ópticos perpendiculares a la línea base, y sus *líneas de exploración* o *líneas epipolares* paralelas a la línea base. Las líneas epipolares son líneas que unen la imagen izquierda y la imagen derecha de un mismo punto. Cualquier punto del espacio tridimensional unido a los dos centros de proyección de las cámaras define un plano, llamado *plano epipolar*. La intersección de un plano epipolar con el plano de proyección de una cámara define una línea epipolar. Para todos los puntos, cuyas proyecciones izquierdas estén contenidas en una misma línea epipolar en la imagen izquierda, sus proyecciones derechas deben estar también contenidas sobre una misma línea epipolar en la imagen derecha, y viceversa. Como se ha dicho anteriormente, el principal problema al realizar visión estereoscópica es encontrar la correspondencia entre los puntos de las imágenes. Al realizar esta correspondencia se debe efectuar una búsqueda en dos dimensiones, tanto en el eje X

como en el eje Y. Cada punto de la imagen derecha al que se quiere emparejar con un punto de la imagen izquierda, se selecciona un conjunto de vecinos (vecindario) en torno al píxel con las mismas coordenadas (sistema de referencia relativo de las cámaras) en la imagen izquierda. Los píxeles que constituyen este vecindario, tienen coordenadas que difieren del píxel de la imagen derecha tanto en la componente X como en la componente Y. Para reducir la complejidad de este proceso se puede llegar a una búsqueda unidimensional, reduciéndose su vecindario a una única dimensión. Para conseguir esta reducción en la complejidad computacional, basta con situar y orientar las cámaras de forma que sólo exista un desplazamiento horizontal entre ellas, los ejes ópticos deben ser paralelos y los ejes de abscisas (eje X) de cada una de las cámaras deben de ser coincidentes como sucede en la Figura 2-2. Al conseguir esta distribución y orientación de las cámaras, éstas sólo tendrán desplazamiento en sentido horizontal, y se dice que las imágenes están alineadas horizontalmente. Bajo esta situación, las líneas epipolares que definen el plano epipolar son coincidentes, consiguiendo simplificar la búsqueda del emparejamiento a recorrer las imágenes por filas. Con esta geometría se obtiene la denominada restricción epipolar, que ayuda a limitar el espacio de búsqueda de correspondencias, de manera que en el sistema de ejes paralelos convencional todos los planos epipolares originan líneas horizontales al cortarse con los planos de las imágenes. En un sistema con la geometría anterior se obtiene un valor de *disparidad d,* para cada par de puntos emparejados P_I (x_I, y_I) y P_D(x_D, y_D) dado por $d = x_I - x_D$. Con el valor de disparidad para cada punto de la imagen se construye un matriz o mapa de disparidad, en el que cada punto de la imagen contiene su valor de disparidad. A partir de este mapa de disparidad se logrará construir el mapa de profundidades, que es el objetivo final de todo sistema estereoscópico.

En la Figura 2-2 se pueden ver todos los elementos comentados. Existen tres sistemas de referencia, uno absoluto y dos relativos a las cámaras. El sistema absoluto, que pertenece a la escena real tridimensional que se está observando, está dado por el origen de coordenadas O y los ejes {X, Y, Z}. El sistema de referencia relativo de la cámara izquierda está dado por el origen de coordenadas O_I y los ejes $\{X_I, Y_I, Z_I\}$, y el

sistema de referencia que falta es el de la cámara derecha, donde su origen de coordenadas es O_D y los ejes $\{X_D, Y_D, Z_D\}$.

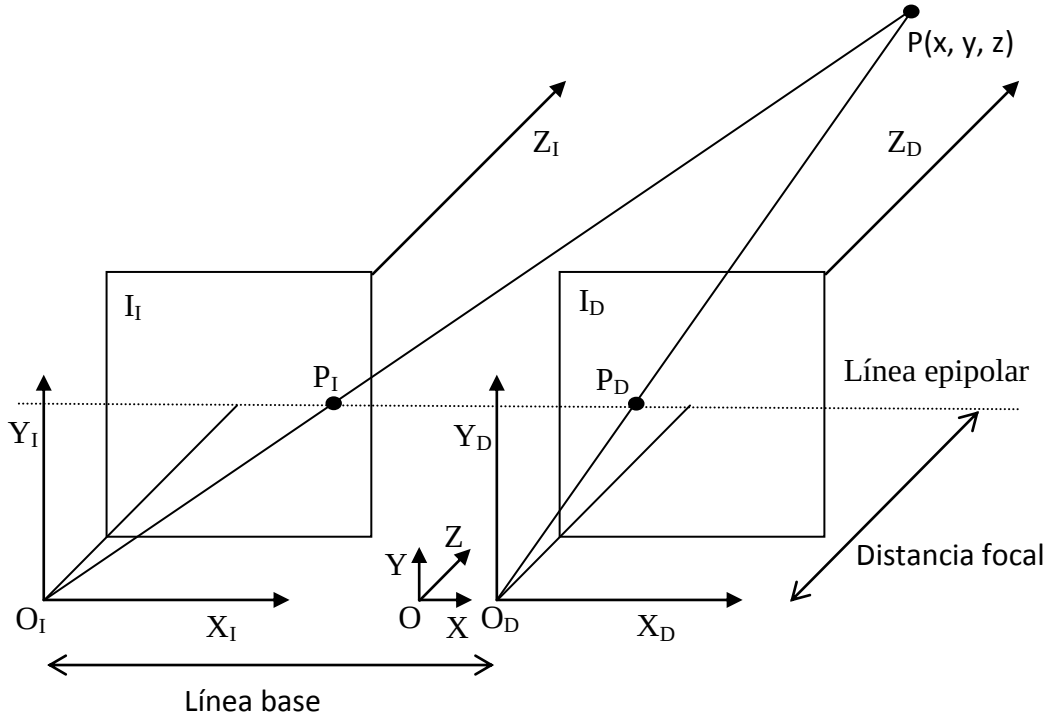


Figura 2-2: Correspondencia de un sistema estero

En ambos sistemas, el de la cámara izquierda y el de la derecha, se hace coincidir el centro de proyección óptico con el origen de coordenadas. Según esta notación, todos los elementos que se refieran a la cámara izquierda y se expresen en su sistema de referencia tendrán el subíndice I, y lo mismo para la cámara derecha pero con el subíndice D. El punto P en la escena tridimensional se proyecta en la imagen izquierda como el punto P_I y en la imagen derecha tiene como proyección P_D. Para obtener la proyección de un punto en una imagen se hace pasar un rayo por su centro óptico y por el propio punto, en la intersección de este rayo con el plano imagen se formará la proyección del punto. Los rayos de proyección PO_I y PO_D definen el plano de proyección del punto de la escena *3-D*, (el plano epipolar). Como se puede comprobar en la figura se han hecho coincidir los ejes X de los dos sistemas relativos, así las imágenes estará en correspondencia y las coordenadas Y de los puntos P_I y P_D

serán idénticas, siendo las líneas epipolares paralelas y consiguiendo simplificar las búsqueda.

En Figura 2-3 se muestra una cámara junto a su sistema de coordenadas. El origen de coordenadas O coincide con el centro de proyección, y el eje Z está superpuesto con el eje óptico del sistema.

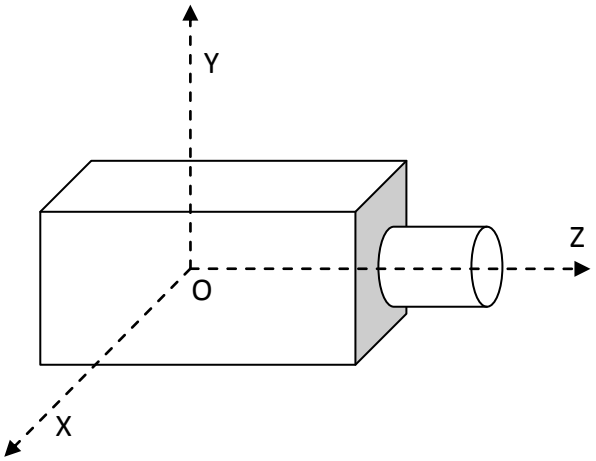


Figura 2-3: Situación del sistema de referencia para una cámara

A pesar de la teoría expuesta anteriormente, en la realidad muy pocas veces las imágenes que forman el par estereoscópico se encuentran alineadas horizontalmente. Además en el proceso de adquisición se introducen deformaciones en las imágenes que degradarán la precisión en la medida de profundidad a menos que sean corregidas. Dos ejemplos de estas deformaciones son la distorsión radial y la distorsión tangencial. Necesitándose en este caso realizar un proceso previo de calibración de cámaras con el fin de corregir dichas anomalías (Pajares y Cruz, 2007).

2.2.3. Extracción de las características

En el paso de extracción de características se obtienen elementos identificativos de la imagen. De estos elementos, a su vez se tienen que extraer algunos atributos, los cuales se utilizarán en el siguiente paso, correspondencia de características. Por lo tanto, este paso está muy ligado al de correspondencia y, como se ha dicho anteriormente, al ser el paso de correspondencia el más importante de todos, primero se suele decidir qué método utilizar al realizar la correspondencia entre

imágenes y según las características que se empleen, éstas serán las que se extraigan de las imágenes. Existen dos clases de técnicas para determinar la correspondencia entre dos imágenes estereoscópicas: las técnicas *basadas en el área* ("area-based") y las técnicas *basadas en las características* ("feature-based").

En las técnicas de estereovisión basadas en el área se busca correlación cruzada entre patrones de intensidad en la vecindad local o vecindario de un píxel en una imagen, con patrones también de intensidad en una vecindad correspondiente a un píxel en la otra imagen del par estereoscópico. Por tanto, las técnicas basadas en el área utilizan la intensidad de los píxeles como característica esencial.

Para las técnicas de estereovisión basadas en las características se toman representaciones simbólicas obtenidas de las imágenes de intensidad en vez de utilizar las intensidades directamente. Algunas de las características cuyo uso está más extendido son: puntos de borde aislados, cadenas de puntos de bordes o regiones delimitadas por bordes.

En los tres tipos de características que se acaban de mencionar se suelen tener en cuenta los puntos de borde para su obtención. De lo anterior se deduce que los puntos de borde utilizados como primitivas son muy importantes en cualquier proceso de visión estereoscópica y en consecuencia, suele ser habitual el hecho de extraer los puntos de borde de las imágenes del par estereoscópico. Una vez que se han extraído los correspondientes puntos de borde algunos métodos utilizan cadenas de puntos de borde, dichas cadenas pueden ser segmentos rectos, segmentos no rectos, cadenas cerradas formando estructuras geométricas con forma definida (elipse, circunferencia, etc.) o estructuras de formas geométricas desconocidas.

Aparte de los bordes, las regiones son otra de las primitivas que pueden utilizarse en el proceso de visión estereoscópica. Una región es una zona de la imagen, que habitualmente está asociada con una determinada superficie en la escena 3-D y delimitada por bordes.

Una vez obtenidas las características y dependiendo del método que se vaya a utilizar, puede ser necesario un paso de segmentación adicional entre la adquisición de

las características y la correspondencia estereoscópica. En este paso se extraería información adicional de las características ya definidas, esta información extra se especifica en una secuencia de propiedades o atributos asociados con las características y se codifica y cuantifica en un vector de atributos, cuyas componentes son precisamente dichas propiedades. Como ejemplo práctico, supóngase un caso en el que las características que se hayan recuperado sean las regiones y de ellas se obtiene información sobre el área, nivel medio de intensidad, coeficiente de textura y elongación. Con esto se obtiene un vector de atributos con cuatro componentes por cada región existente, este vector tiene como primera componente el área de la región seleccionada (x_1), como segunda componente el valor medio de las intensidades que constituyen el región (x_2), como tercera componente el coeficiente de textura de la región (x_3) y como cuarto y último componente está la elongación de la región (x_4). Finalmente se tendrían tantos vectores como regiones, por cada región *r* existiría un vector X_r asociado a la región y tendría el aspecto $X_r = \{x_1, x_2, x_3, x_4\}$. Al estar toda esta información almacenada en vectores, será independiente la cantidad o significado de los atributos que se utilicen, los métodos que tengan como entrada vectores de atributos realizarán las mismas operaciones sin considerar qué tipo de característica se ha utilizado.

2.2.4. Correspondencia de características

En el paso de la correspondencia de imágenes, bien utilizando píxeles o características, se debe determinar para un punto del espacio tridimensional, cuál es su proyección en cada imagen del par estereoscópico. Al comienzo del proceso de correspondencia ya se tienen los vectores con los atributos de las características consideradas. Con estos vectores, comparando los valores que toman sus atributos se debe establecer una correspondencia local entre características. Esta correspondencia se determina mediante alguna métrica que proporcione cuál es el grado de similitud para dos vectores de atributos, como por ejemplo la distancia Euclídea, la distancia de Mahalanobis u otras distancias. Tras realizar esta correspondencia local, se debe comprobar su consistencia, para lo cual se comienza otro proceso de correspondencia, pero en este caso de naturaleza global. Ambos procesos de correspondencia utilizan propiedades de la realidad física, que son formuladas en términos de restricciones. Las

restricciones utilizadas se enuncian a continuación (Pajares y Cruz, 2007; Scharstein y Szeliski, 2002):

- *Epipolar*: las imágenes de una misma entidad 3D deben proyectarse sobre la misma línea epipolar. Esta restricción se deriva de la geometría de las cámaras y requiere que las cámaras estén alineadas. Un método para lograr el alineamiento se puede encontrar en Reid y Beardsley (1996).
- *Semejanza*: las dos imágenes de la misma entidad 3D deben tener propiedades o atributos similares.
- *Unicidad*: para cada característica en una imagen debe haber una única característica en la otra imagen, salvo que se produzca una oclusión y no haya correspondencia de alguna característica.
- *Orden posicional*: dadas dos características en una determinada imagen, por ejemplo la izquierda, situada una a la derecha de la otra, esta restricción supone que este mismo orden se mantiene en la imagen derecha para sus respectivas características homólogas. Debido a la geometría también se debe verificar que, para una misma característica el valor de su coordenada *x* en la imagen derecha debe ser menor que su coordenada x en la imagen izquierda. Dado el par de imágenes estereoscópicas de la Figura 2-4, supongamos que el punto ***a*** y la estrella ***b*** de la imagen izquierda se corresponden con el punto ***c*** y la estrella ***d*** de la imagen derecha, como se tiene que el punto está más a la izquierda que la estrella en la imagen izquierda, $x_a < x_b$, también debe de estarlo en la imagen derecha, $x_c < x_d$. Además, la coordenada horizontal del punto y de la estrella en la imagen derecha debe de ser que la coordenada horizontal en la imagen izquierda, $x_c < x_a$ y $x_d < x_b$.

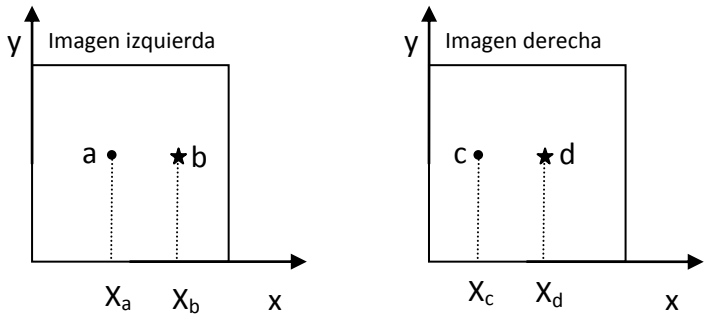


Figura 2-4: Restricción sobre el orden posicional

- *Continuidad de la disparidad*: asume que las variaciones de disparidad en la imagen son generalmente suaves, es decir que si consideramos un mapa de disparidad éste se presenta continuo salvo en unas pocas discontinuidades. Este principio también aparece bajo distintas formas y a veces con alguna pequeña variación, tal es el caso de *Disparidad Diferencial Mínima* en Medioni y Nevatia (1985) o Pajares *et ál.* (1998) o *Continuidad "figural"* en Pollard *et ál*. (1981).
- *Relaciones estructurales*: supone que los objetos están formados por aristas, vértices o superficies con una cierta estructura y una disposición geométrica entre dichos elementos.

Las restricciones estereoscópicas anteriormente descritas, que se aplican sobre las imágenes analizadas en este trabajo dada la naturaleza de las mismas son las siguientes: epipolar, semejanza, unicidad y continuidad de la disparidad. Por otra parte se ha decidido utilizar la correspondencia píxel a píxel, por lo que en resumen el proceso de correspondencia llevado a cabo se concreta como sigue tomando como referencia la figura 2.5.

La restricción epipolar permite restringir el espacio de búsqueda para la correspondencia entre píxeles, ya que la geometría del sistema así lo determina.

La restricción de semejanza permite establecer medidas de similitud entre pares de píxeles procedentes de la imagen izquierda con píxeles de la imagen derecha y viceversa, de derecha a izquierda. De esta forma, para cada píxel de una de las imágenes se dispone de un conjunto de píxeles en la otra imagen, que constituyen una

lista de posibles candidatos. El objetivo final es elegir de entre los candidatos de la lista uno de ellos como el preferido. Llegados a este punto, todavía restan por aplicar las restricciones de continuidad de la disparidad y unicidad. Dependiendo del orden en el que se apliquen a partir de este momento, caben dos alternativas de diseño posibles para completar el proceso de correspondencia estereoscópica, a saber:

- aplicar unicidad seguida de continuidad, Figura 2-5 , rama izquierda
- aplicar continuidad y posteriormente unicidad, Figura 2-5 ,rama derecha

En el primer caso, una vez que se dispone de un conjunto de medidas de similitud entre píxeles de la imagen izquierda con píxeles de la derecha, se procede a tomar una decisión sobre cuál es la correspondencia considerada como verdadera. Esto se puede llevar a cabo por cualquier mecanismo de decisión, de forma que mediante la aplicación de la restricción de unicidad, se elige un único candidato de entre todos los posibles. De esta forma se obtiene un mapa de disparidad inicial. Lo más probable es que dicho mapa contenga errores, incluyendo falsos positivos (Rohith *et ál.*, 2008). La idea principal a partir de aquí consiste en aplicar el concepto de continuidad de la disparidad, de forma que siguiendo los principios de la Gestalt (Wang, 2005), se asume que las disparidades de píxeles vecinos son similares excepto en determinadas zonas de discontinuidad. En esta línea se han utilizado distintas técnicas tales como filtros bilaterales (Ansar, Castano y Matthies, 2004), ajustes de planos sobre los mapas de disparidad (Klaus, Sorman y Karner, 2006) o filtros estadísticos como media, mediana o moda (Lankton, 2010). En cualquier caso, todas estas técnicas aplican la restricción de continuidad de la disparidad, tras lo cual se obtiene un nuevo mapa de disparidad refinado, donde determinados valores de disparidad espurios o erróneos se consiguen eliminar.

En el segundo caso, tras la aplicación de la restricción de semejanza, no se toma todavía una decisión encaminada a seleccionar emparejamientos correctos, antes bien, se opta por aplicar la restricción de continuidad sobre la base de que dado un píxel en la imagen izquierda y su homólogo en la derecha, los píxeles vecinos de ambos, en sendas imágenes, deben ser también homólogos entre sí, asumiendo que la disparidad es continua en dichas vecindades. Estos procesos suelen ser de

optimización basados en la minimización de una función de energía (Banno e Ikeuchi, 2009; Ruichek y Postaire, 1996), de suerte que una vez finalizado el proceso de optimización, los píxeles han modificado los valores iniciales de similitud, adquiriendo nuevos valores, en base a los cuales y por aplicación de la restricción de unicidad se obtiene el mapa de disparidad. Para ello, como en el caso del mapa inicial, se aplica algún proceso de decisión.

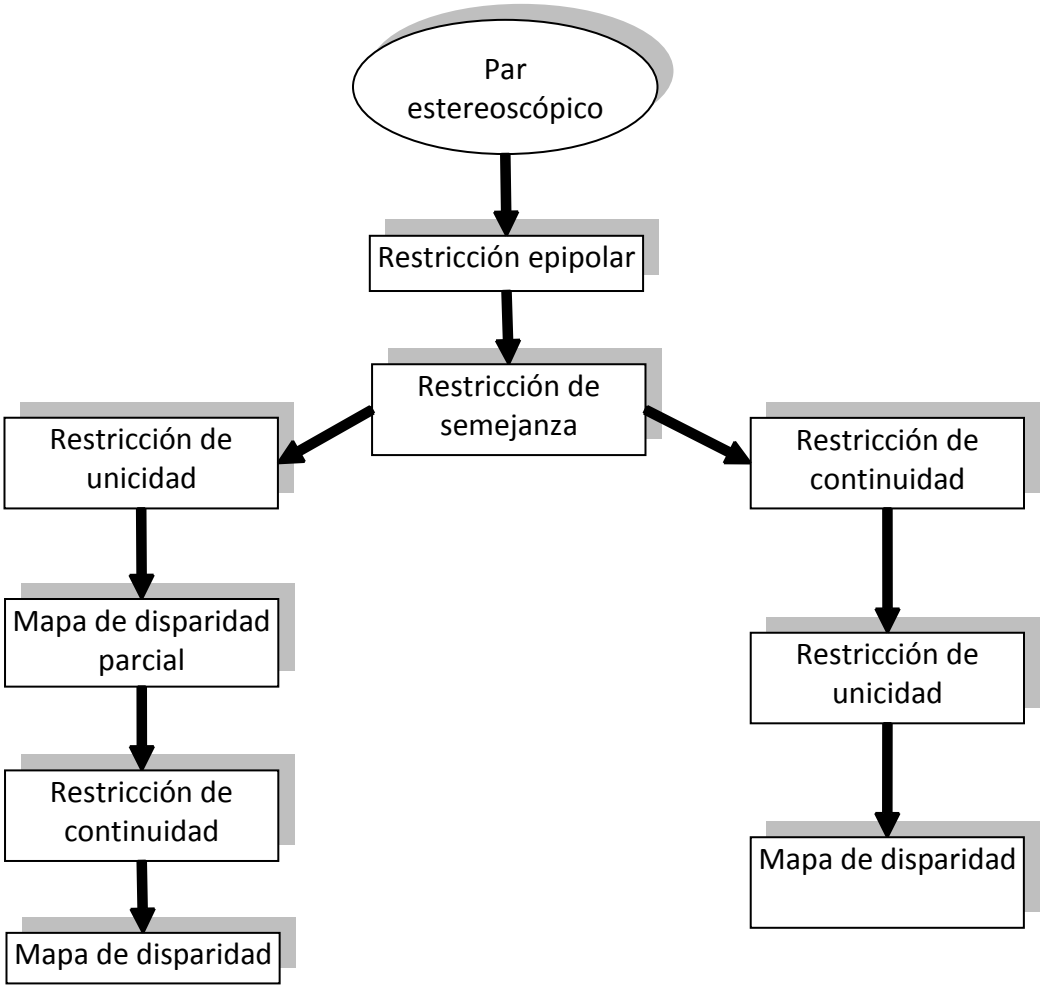


Figura 2-5: Distintos enfoques seguidos en la etapa de correlación, según el orden de aplicación de las restricciones de unicidad y continuidad

En el estudio llevado a cabo en este trabajo se ha decido seguir el primer enfoque con el fin de tratar de mejorar el mapa de disparidad en la medida de lo posible, de esta forma, se aplica primeramente la restricción de unicidad y posteriormente la de continuidad. Se ha optado por este debido a que tras la

restricción de semejanza todavía existen muchos posibles candidatos entre los que elegir, y aplicar la restricción de continuidad trae consigo dos problemas importantes. El primer problema es la cantidad de información que se debe almacenar y tratar, ya que se deber aplicar la restricción de continuidad a todos los posibles candidatos. Y el segundo problema es que entre los posibles candidatos algunos incumplen fuertemente la restricción de continuidad, y si se aplica la restricción de continuidad a estos elementos se podría distorsionar la realidad, llegando a seleccionar a estos como verdaderas correspondencias cuando en realidad no lo son.

Así a modo de resumen se ha decidido escoger el primer enfoque, en el que primero se aplica la restricción de unicidad, obteniendo un mapa de disparidad temporal, sobre el que después se aplica la restricción de continuidad para obtener el mapa de disparidad final.

Como ya se ha expresado en varias ocasiones, la correspondencia es el paso más complejo del proceso de la visión estereoscópica. Al ser el paso más complicado es también el aspecto más estudiado y repetitivo en la literatura de la estereovisión. La dificultad para establecer la correspondencia entre puntos o características de un par de imágenes estereoscópicas proviene de la naturaleza del propio sistema. Las imágenes que forman el par estereoscópico son tomadas por un sistema de visión estéreo, donde la diferencia que existe entre la imagen izquierda y derecha es el ángulo o posición con las que fueron realizadas. Teóricamente esta es casi la única diferencia que se produciría, pero en la práctica las condiciones de iluminación pueden ser ligeramente diferentes o incluso existe la posibilidad de que aparezcan reflejos en una imagen y que la otra imagen esté ausente de ellos. A su vez se suele producir un fenómeno adicional que complica las cosas de manera importante, son las oclusiones, donde una característica de una imagen puede ocultarse en la otra del par estereoscópico. Además, los sistemas convencionales utilizan dos o más cámaras, dichas cámaras aún siendo de las mismas características técnicas presentan diferencias intrínsecas debido al distinto comportamiento de los componentes ópticos y electrónicos.

Debido a los anteriores factores, un mismo punto y su vecindad en la escena 3-D puede proyectarse en cada una de las imágenes con diferentes valores de intensidad o incluso estar presente en una imagen y ausente en la otra imagen si se produce una oclusión. Esto es por lo que el proceso de correspondencia estereoscópica es el paso más difícil y complicado dentro del proceso de visión estereoscópica.

Se mencionó en el paso relativo a la "extracción de las características" que existen dos tipos de técnicas para realizar la correspondencia estereoscópica: las técnicas basadas en el área y las técnicas basadas en las características. Ambas técnicas tienen sus ventajas y desventajas, y dependiendo de cuáles sean las restricciones a las que nos enfrentamos se debe elegir una u otra técnica. Habrá ocasiones en las que la técnica a utilizar venga más o menos impuesta (por ejemplo, si en el entorno a estudiar se dispone de unos bordes perfectamente definidos, utilizando una técnica basada en las características se podría realizar la correspondencia con relativa facilidad) y, sin embargo, en otras ocasiones será difícil decantarse por unas u otras, o incluso se puede realizar una mezcla de las dos. Las principales ventajas e inconvenientes de los dos tipos de técnicas se pueden encontrar en Pajares y Cruz (2007) detallándose a continuación por su importancia.

La principal ventaja de las técnicas basadas en el área es que producen un mapa de disparidad denso ya que la correspondencia se realiza píxel a píxel y se obtiene un valor de disparidad para cada píxel. No obstante estos métodos conllevan una serie de inconvenientes, a saber:

- Usan valores de intensidad en cada píxel directamente, por lo que presentan una elevada sensibilidad a distorsiones debidas a la variación del punto de vista (perspectiva) así como a cambios en la intensidad absoluta (contraste e iluminación).
- La presencia de bordes en las ventanas de correlación originan falsas correspondencias, puesto que hacen que las superficies presenten discontinuidades o que en una imagen se haya ocultado un borde con respecto a la otra.

- Están fuertemente ligadas a la restricción de epipolaridad, mencionada anteriormente.

Frente a las técnicas basadas en el área están las técnicas basadas en las características, éstas utilizan las primitivas obtenidas en el paso de extracción de características para hallar la correspondencia. Las principales ventajas de estas técnicas son:

- mayor estabilidad frente a cambios en el contraste e iluminación ambiente,
- permiten realizar comparaciones entre los atributos o propiedades de las características,
- mayor rapidez que los métodos basados en el área puesto que hay menos puntos (características) a considerar, aunque requieren un tiempo de procesamiento previo, que consume tiempo
- la correspondencia es más exacta puesto que los bordes pueden ser localizados con mayor precisión,
- son menos sensibles a variaciones fotométricas porque representan propiedades geométricas de la escena,
- no poseen una fuerte dependencia de la restricción epipolar,
- representan aspectos más abstractos de la escena, siendo menos sensibles al ruido, ya que ha habido un procesado previo.
- enfocan su interés hacia lugares de la escena donde el contenido de información es máximo,
- se han utilizado en entornos industriales por su robustez (Lane *et ál*. 1994).

A pesar de esta serie de ventajas, las técnicas basadas en las características tienen dos principales inconvenientes. Uno es que, son muy dependientes de las primitivas que se escojan, pudiéndose producir resultados malos o poco fiables si las primitivas elegidas no son acertadas o poco apropiadas para el tipo de imagen. Por ejemplo, en una escena donde haya pocos bordes y los que existen estén poco definidos no sería recomendable seleccionar delimitadores de regiones como primitivas. El segundo inconveniente es que en comparación con los métodos basados en el área, los métodos basados en las características producen mapas de profundidad

poco densos, por lo que muy probablemente se necesitará un paso posterior para dotar de una mayor densidad al mapa. Este paso es el último de los que se compone el proceso de visión estereoscópica, antes de este último paso se debe obtener un mapa de profundidades, ya que hasta el momento se está trabajando con disparidades. Por tanto, el siguiente paso será transformar el mapa de disparidades, que acabamos de obtener en el paso de correspondencia, en un mapa de profundidades.

2.2.5. Determinación de la distancia

Una vez que se ha hecho corresponder los elementos que aparecen en la imagen izquierda con los elementos en la imagen derecha, se dan las condiciones necesarias para continuar el proceso de visión estéreo y hallar la distancia a la que se encuentran los objetos que aparecen en la escena. Una vez que se ha efectuado el proceso de correspondencia de forma precisa, la determinación de la profundidad es un proceso relativamente sencillo, reduciéndose a una simple triangulación. Sin embargo en algunas ocasiones, cuando se intenta hallar la distancia a la que se encuentra una característica se presentan algunas dificultades debidas a una falta de precisión o una escasa fiabilidad cuando se intentó encontrar la correspondencia.

Debido a las restricciones de epipolaridad, las proyecciones de un objeto real en cada una de las imágenes solamente se diferenciarán en la coordenada x de sus sistemas de referencia relativos, ya que la coordenada y será idéntica. De forma que dos elementos de las imágenes, que representan al mismo objeto, solo tendrán desplazamiento horizontal. Exactamente este desplazamiento es lo que hemos llamado disparidad. Para cada característica que hayamos encontrado su correspondencia con otra característica de la otra imagen del par estereoscópico, tendremos un valor de disparidad. Todos estos valores están almacenados en la matriz o mapa de disparidad.

Considerando una relación geométrica de semejanza de triángulos, las coordenadas del punto de la escena $P(X, Y, Z)$ pueden deducirse fácilmente sin más que observar la Figura 2-6, obteniendo los resultados dados por la ecuación (2-1). La Figura 2-6 es el mismo sistema y situación que se daba en el Figura 2-2, pero observado desde arriba. Se ha utilizado esta perspectiva superior porque se aprecia

mejor las distancias utilizadas en el proceso de triangulación. En este caso se denota la distancia de la línea base con la letra *b* y la distancia focal, idéntica en las dos cámaras, con la letra *f*. En la figura, P es un punto de la escena con coordenadas (x, y, z) según el sistema de coordenadas del mundo real; P_I es la proyección del punto P en la imagen izquierda y tiene por coordenadas (x_I,y_I), según el sistema de coordenadas relativo de la cámara izquierda; y P_D es la proyección del punto P en la imagen derecha, que tiene por coordenadas (x_D,y_D), según el sistema de coordenadas relativo de la cámara derecha. Con la letra d se representa la disparidad del punto, que es el desplazamiento horizontal que se producen, $d = x_I - x_D$.

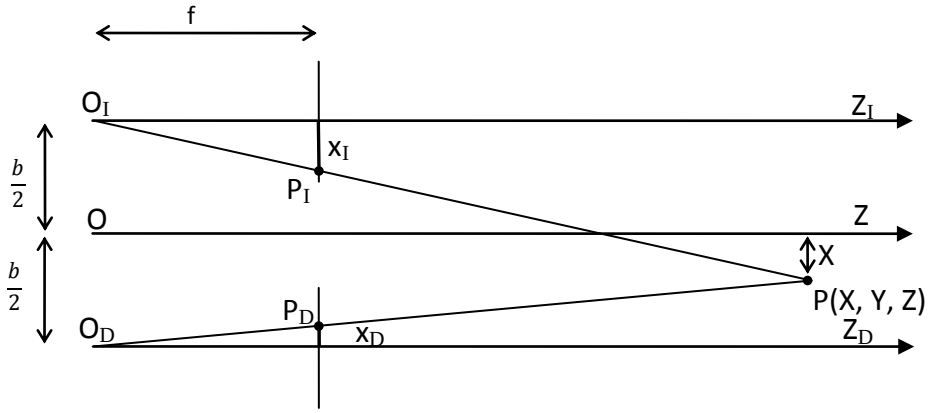


Figura 2-6: Geometría de dos cámaras en estéreo con ejes ópticos paralelos desde una perspectiva superior. El valor de la línea base está determinado por el valor b y f es la distancia focal de las dos cámaras. P es un punto de la realidad y P_I y P_D son sus proyecciones en la imagen izquierda y derecha respectivamente.

Se deduce a partir de la ecuación (2-1) que cuando se utiliza esta geometría, la profundidad *Z*, es inversamente proporcional a la disparidad de la imagen y para una profundidad dada, a mayor *b* mayor *d*, lo que sugiere que la línea base puede incrementarse para mejorar la exactitud de la profundidad medida, pero ello lleva consigo el hecho de que ahora ambas imágenes tienen menos características comunes, es decir menos bordes o regiones procedentes de los objetos de la escena, debido a las desapariciones y oclusiones de las imágenes de dichos objetos.

$$\text{Imágen izquierda:}\frac{b/2+x}{z}=\frac{x_I}{f} \qquad x_I=\frac{f}{z}\left(x+\frac{b}{2}\right)$$
$$\Rightarrow \qquad \text{(2-1)}$$
$$\text{Imágen derecha:}\frac{b/2-x}{z}=\frac{x_D}{f} \qquad x_D=\frac{f}{z}\left(x-\frac{b}{2}\right)$$
$$d=x_I - x_D=\frac{f\cdot b}{z}\Rightarrow z=\frac{f\cdot b}{d}$$

2.2.6. Interpolación

Este paso no resulta siempre de aplicación, se utiliza bajo el enfoque basado en las características ya que es cuando la información sobre distancias puede ser insuficiente. Normalmente las aplicaciones requieren de mapas de profundidad más o menos densos. Que un mapa de profundidad sea denso significa que hay mucha información sobre la distancia a la que se encuentran los elementos de la escena por unidad de superficie, frente a un mapa de superficie disperso en el que la información que se dispone por unidad de superficie es menor. La densidad de un mapa es un valor que se puede cuantificar más o menos, pero que se considere lo suficientemente denso o no depende únicamente de la aplicación. Por ejemplo un mismo mapa que sea lo suficientemente denso para una aplicación puede que no sirva para otra por ser demasiado disperso para ella, o que teniendo dos mapas de los cuales uno tenga una mayor densidad que el otro, puede que el que tenga una mayor densidad no sirva para una aplicación por no ser lo suficientemente denso y, sin embargo, el otro mapa que es más disperso sí que sirva para otra aplicación porque esta tenga unos requisitos menores sobre profundidad. Sin menoscabo de lo que se acaba de comentar, un mapa de disparidad, donde para cada píxel de la imagen capturada se disponga de un valor de la distancia a la que se encuentra ese punto de la cámara en la escena tridimensional, se puede considerar muy denso, aunque también existen técnicas con las conseguir una mayor resolución para el mapa de profundidad, siendo la precisión a nivel de subpíxel (Gehrig y Franke, 2007).

Con los métodos de correspondencia basados en características se suele llegar a mapas de disparidad dispersos, porque las características que tienen las imágenes

suelen estar diseminadas y distribuidas irregularmente. Los métodos de correspondencia basados en el área son más apropiados que los métodos de correlación basados en las características si se quiere obtener un mapa de profundidades denso, aunque la información de los métodos basados en las características sea más fiable. La información de los métodos basados en el área suele ser menos fiable por problemas tales como la existencia de zonas donde el brillo es muy uniforme y no se puede realizar una correspondencia con certeza. Debido a estas zonas con falta de fiabilidad también se puede decidir utilizar interpolación a pesar de utilizar métodos basados en el área.

Uno de los modos más sencillos de solucionar el problema de la interpolación es interpretar el mapa de disparidad como un muestreo de una función de profundidad continua, y utilizar un método de interpolación tradicional para hallar la función continua que la aproxime. Así se obtendría una función continua con la que obtener la profundidad para cualquier punto de espacio tridimensional, consiguiendo transformar un mapa de profundidad disperso en denso. Si el mapa de profundidad disperso capturaba los principales cambios en la profundidad, el mapa que se alcanza será válido. Métodos que se pueden utilizar para interpolar son: la interpolación de Lagrange, la interpolación de Hermite, interpolación mediante Splines, o mediante wavelets.

También se puede resolver el problema de la interpolación por otro camino, utilizando modelos geométricos calculados previamente e intentando realizar un ajuste con la matriz dispersa de profundidades. Antes de intentar el ajuste entre el modelo geométrico y la matriz de profundidades, se lleva a cabo un proceso de *clustering*. Con el clustering se obtienen los puntos de la escena tridimensional que pertenecen a elementos destacados. Después de que se haya dividido la escena en clústeres, se comprueba para cada clúster a qué modelo de los existentes se ajusta mejor. Trabajos en los que se ha utilizado el reflejo especular que se produce en los objetos para determinar su superficie y poder así estimar sus tres dimensiones se tienen en Solem *et ál.* (2007) y en Jin *et ál.* (2001).

2.3.Revisión de técnicas para la correspondencia estereoscópica

Como se ha explicado en la sección 2.2.4 existen dos modos de abordar el problema de la correspondencia en visión estereoscópica, estos modos hacen referencia a los dos tipos de técnicas empleadas para emparejar características de una imagen con características de la otra imagen del par estéreo. Se tienen las técnicas basadas en el área y las técnicas basadas en las características. En esta sección se hará un recorrido por las técnicas actuales en correspondencia estereoscópica, realizando una distinción entre los dos tipos de técnicas. Si bien, la tendencia actual es aprovechar las ventajas de ambas y utilizarlas conjuntamente. Brown *et al. (*2003) realiza una clasificación de métodos atendiendo al criterio de correspondencia local o global. Esta clasificación de los métodos utilizados para hallar la correspondencia estereográfica no aparece por primera vez en este artículo, sino que ya se hablaba de métodos globales con anterioridad (Pajares *et al*., 1998). Estos métodos deben su nombre por el tipo de restricciones que utilizan. Básicamente las primeras se refieren a la aplicación de las restricciones a nivel de característica o de un conjunto restringido de características mientras que las segundas las aplican sobre las características que aparecen en el conjunto de las imágenes.

2.3.1. Técnicas basadas en el área

Las técnicas basadas en el área utilizan patrones de intensidad en la vecindad local de un píxel en una imagen con patrones también de intensidad en una vecindad homóloga de un píxel en la otra imagen del par estereoscópico (Cochran y Medioni, 1992; Scharstein y Szeliski, 2002).

Uno de los métodos más sencillos que se pueden aplicar a la hora de realizar correspondencia basada en el área es, a partir de un punto o píxel de una imagen, por ejemplo de la derecha, seguir la línea epipolar que pasa por él y realizar una búsqueda en la otra imagen, la izquierda, para encontrar el píxel que tenga la intensidad más parecida al de partida. En este método se están aplicando dos restricciones: la de epipolaridad, ya que se sigue la línea epipolar que pasa por el píxel para determinar su homólogo en la otra imagen del par y la de semejanza, para determinar el píxel homólogo en la imagen izquierda se comprueba qué píxeles tienen intensidad

semejante. Este método no garantiza resultados fiables, ya que pueden existir píxeles con la misma intensidad y el método no tendría modo alguno para discernir cuál de todos los píxeles es el que determina una verdadera correspondencia. Por tanto, se debe introducir algún procedimiento para que la técnica sea lo más robusta posible.

Aplicando más restricciones se puede aumentar la fiabilidad del método. En efecto, se puede aplicar además de la restricción epipolar y de semejanza, la de orden posicional de forma que teniendo en cuenta la geometría del sistema, la búsqueda de correspondencias en la imagen izquierda se realice siempre en las posiciones con valores en la coordenada $x_i > x_d$, donde x_d y x_i son las correspondientes coordenadas en el eje *x* de las imágenes derecha e izquierda, respectivamente. En este sentido también conviene considerar no sólo el punto bajo correspondencia sino un entorno de vecindad del mismo.

Con el criterio anterior se establece un entorno de vecindad definiendo una ventana de dimensión *MxN* alrededor del punto de origen en la imagen derecha. A las técnicas que utilizan estas ventanas o bloques, en ocasiones se les denomina en la literatura *"técnicas basadas en correspondencia de bloques"*. Normalmente se define una vecindad cuadrada, donde el número de columnas y filas que constituyen la ventana es el mismo. En cada punto de la imagen izquierda con coordenadas $x_i > x_d$ y situados sobre la misma línea epipolar se define también una ventana de las mismas dimensiones que la anterior. Gracias a estas ventanas, ahora no se tiene en cuenta solamente el punto de interés, sino que se valora tanto éste, sobre el que se encuentra centrada la ventana, como su vecindad. Gracias al uso de esta ventana se pueden solventar en ciertas ocasiones los problemas de tener píxeles candidatos con semejante niveles de intensidad en la imagen izquierda, bastaría con observar las intensidades de los píxeles pertenecientes a la vecindad definida por la ventana y comprobar si también son semejantes.

Para determinar la similitud se emplean medidas estadísticas o distancias métricas entre píxeles. Un ejemplo de medida estadística sería la correlación (Pajares y Cruz, 2007). Para cada píxel de la imagen derecha se halla la desviación típica y la media de la ventana asociada a este píxel, se recorren los píxeles de la imagen

izquierda que caen sobre la línea epipolar y que satisfacen la restricción de orden posicional, hallándose también sus desviaciones típicas y sus medias, así como la covarianza entre estos píxeles y el supuestamente homólogo de la imagen derecha. Tras recorrer todos los píxeles de la imagen izquierda que cumplen las restricciones y haber obtenido los valores estadísticos, se obtienen los coeficientes de correlación C. Con todos los coeficientes de correlación se determina cuál es el que tiene un valor más cercano a uno, y el píxel al que corresponda este coeficiente será el que tendrá una mayor semejanza con el píxel de la imagen derecha, y por tanto será con el que se establezca la correspondencia. La normalización de este método, tanto en la media como en la varianza/desviación típica, hace de él un método relativamente insensible a variaciones radiométricas.

De cara a obtener una función de evaluación simple, pero manteniendo el uso de ventanas rectangulares de dimensión $N \times M$, Shirai (1987) propone usar la suma de las diferencias de las intensidades para los puntos en la ventana como medida de similitud.

$$D = \sum_{i=1}^{M} \sum_{j=1}^{N} \left(I_I(i,j) - I_D(i,j) \right)^2 \qquad \text{(2-2)}$$

En la ecuación (2-2) $I_k(i,j)$ representa la intensidad del píxel (i,j) para la imagen k, donde k = I se refiere a la imagen izquierda y si k = D a la derecha. El valor de D es inversamente proporcional a la semejanza, por tanto la mayor semejanza entre píxeles se corresponde con el mínimo valor de D. Esta métrica también puede ser normalizada para evitar que las transformaciones radiométricas afecten al resultado, su normalización se efectúa tal y como se expresa en la siguiente ecuación:

$$D = \sum_{i=1}^{M} \sum_{j=1}^{N} \left(\frac{\left(I_I(i,j) - \overline{I_I} \right)}{\sigma_I} - \frac{\left(\left(I_D(i,j) - \overline{I_D} \right) \right)}{\sigma_D} \right)^2 \qquad \text{(2-3)}$$

donde $\bar{I}_D$ y $\bar{I}_I$ son la media de los valores de intensiadad de la ventana derecha e izquierda respectivamente, y σ_D y σ_I son las desviaciones estándar de los valores de intensidad de la ventana derecha e izquierda respectivamente.

Como los valores de D se obtienen secuencialmente y aumenta de forma monótona, se puede detener el cálculo cuando el resultado parcial del sumatorio sea suficientemente grande. Esta idea fue propuesta inicialmente por Barnea y Silverman (1972) y el algoritmo se llamó *SSDA (Sequential Similarity Detection Algorithm)*. La función de evaluación también se simplificó, sustituyéndose el cuadrado por el valor absoluto, con el fin de reducir el coste computacional.

$$D_2 = \sum_{i=1}^{M}\sum_{j=1}^{N}\left|I_I(i,j) - I_D(i,j)\right| \qquad (2\text{-}4)$$

Estos métodos tienen algunos problemas relacionados con el tamaño de la ventana y el valor umbral a partir del cual determinar si existe correspondencia o no. Si la ventana es lo suficientemente grande se pueden atenuar los problemas de ruido, ya que si aparece un píxel erróneo, al disponer de más información se amortigua su acción en el resultado global. Pero al aumentar el tamaño de la ventana también se incrementa el coste computacional al tener que realizar más operaciones. En Shirai (1987) se propone un algoritmo para resolver estas cuestiones, comienza por una ventana de dimensiones pequeñas y si se produce claramente un mínimo por debajo de un umbral fijado, entonces la correspondencia se acepta, si es mayor que un segundo umbral prefijado la correspondencia se rechaza y si la decisión no es clara se realiza el proceso con una ventana de mayores dimensiones. En Aschwanden y Guggenbuhl (1993) se realiza un extenso estudio comparativo entre las métricas anteriores.

Otro método alternativo para hallar la correspondencia que aplica transformaciones no paramétricas locales antes de realizar la correspondencia es el propuesto por Zabih y Woodfill (1994). La transformación para reducir la sensibilidad a las modificaciones radiométricas que se utiliza es la conocida en terminología inglesa

como *rank*. La transformada *rank* para una determinada ventana se define como el número de píxeles de la ventana para los cuales su intensidad es menor que la intensidad del píxel central. Los valores resultantes tras aplicar esta transformación están basados en una ordenación relativa de la intensidad de los píxeles en vez de en las intensidades mismas. Una vez aplicada la transformación rank, se aplican una de las métricas anteriores para establecer la correspondencia.

En la Figura 2-7 se tiene una ventana de dimensión cinco, donde la intensidad del píxel central es de 27. Tras aplicar la transformada rank se resume la información de la ventana en el dígito 5, esto significa que en un entorno de vecindad de cinco para el píxel central, existen cinco píxeles con un valor de intensidad menor que 27 (intensidad del píxel central).

$$\begin{matrix} 81 & 47 & 90 & 58 & 12 \\ 70 & 91 & 34 & 63 & 24 \\ 09 & 75 & 27 & 85 & 54 \\ 69 & 95 & 75 & 96 & 49 \\ 15 & 76 & 97 & 06 & 95 \end{matrix} \xrightarrow{transformada \quad rank} 5$$

Figura 2-7: Transformada rank para una ventana de dimensión cinco

Con la transformación Rank se reduce la sensibilidad a las distorsiones radiométricas, pero también disminuye la capacidad discriminatoria para establecer la correspondencia ya que se pierde información. La ordenación relativa de los píxeles que caen dentro de la vecindad se codificada en un único valor. Zabih y Woodfill (1994) proponen una modificación del método anterior que preserva la distribución espacial de los vecinos, esta variación de la transformada rank la denominan transformada *census*. Esta nueva transformada codifica la información en una cadena de bits en vez de en un único valor. En la Figura 2-8 se puede ver un ejemplo de cómo se aplica la transformada census a un píxel mediante una ventana de dimensión cinco. El píxel al que se le aplica la transformada se encuentra situado en el centro de la ventana con un valor de intensidad de 27. La ventana se recorre desde el extremo superior izquierdo hasta el extremo inferior derecho, procesando la matriz que forma la ventana por filas. Si el valor de intensidad del píxel que se está recorriendo es menor que el valor de intensidad del píxel central, se añade un cero a la cadena y si el valor es menor se añade un uno. Para este ejemplo, donde el valor central es de 27 se tienen cinco

píxeles con un valor inferior, 12, 24, 09, 15 y 06, los cuales participan con un valor de uno en las posiciones 5, 10, 11, 20 y 23 de la cadena.

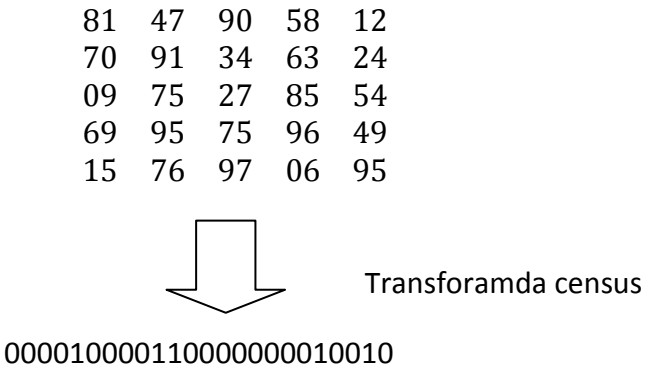


Figura 2-8: Transformada census para una ventana de dimensión cinco

Tras realizar la transformada *census* se puede determinar la correspondencia de un píxel mediante el cómputo de la distancia de Hamming. La distancia de Hamming se realiza entre cada una las cadenas de los píxeles candidatos halladas en una imagen, por ejemplo la izquierda, y la cadena del píxel del que se busca su homólogo en la otra imagen del par, la imagen derecha. El inconveniente de esta transformada es el incremento de información por cada píxel, y el consiguiente aumento en el coste computacional. En Banks y Corke (2001) se realiza un análisis comparativo del rendimiento de la correspondencia alcanzada mediante las transformadas rank y census frente a otros métodos como el de correlación.

Otras técnicas para hallar la matriz de disparidad utilizan una función de energía. En Alagöz (2008) se ha definido una función de energía que toma como argumentos la posición de un píxel en la imagen derecha y un valor de disparidad, y proporciona como salida un valor de energía. Este valor de energía representa la fuerza con la que están en correspondencia el píxel de la derecha, para el que se han introducido sus coordenadas en la imagen, con el píxel de la imagen izquierda resultante de aplicar a este píxel derecho el valor de disparidad de entrada. De forma semejante a como sucede en la realidad, cuanto menor sea el valor energético de esta vinculación, más estable será la unión y más acertada será la correspondencia. Así, el objetivo será obtener el valor de disparidad que proporciona el valor de energía mínimo para cada píxel de la imagen derecha. Esta función trabaja también con

bloques o ventanas, ya que así se obtienen unos resultados más fiables. Finalmente, tras obtener una matriz con las disparidades asociadas a cada píxel de la imagen derecha, se aplica el filtro de la media aritmética varias veces para suavizar la matriz de disparidades. El filtrado de la media (Chan, Osher y Shen, 2001) tiene como objetivo eliminar cambios muy bruscos, los cuales muy probablemente se deban a un error en la correspondencia.

Existen métodos basados en la segmentación de las imágenes que proporcionan unos resultados aceptables (Bleyer y Gelautz, 2005; Deng *et ál.*, 2005; Hong y Chen, 2004; Tao, Sawhney y Kumar, 2001). Estos métodos se basan en la suposición de que la estructura de la escena puede ser aproximada por un conjunto de planos que no se solapan y donde cada plano es coincidente con por lo menos una región de color homogéneo. Los métodos basados en segmentación suelen estar constituidos de cuatro pasos: 1) se localizan las regiones de color homogéneo aplicando un método de segmentación por color; 2) se usa un método de correspondencia basado en el área para determinar la disparidad de los puntos de forma fiable; 3) se aplica alguna técnica para obtener planos de disparidad, que son considerados como un conjunto de etiquetas; 4) se realiza una asignación de planos de disparidad óptima (etiquetado óptimo). En Klaus *et al*. (2006) se propone un método basado en la segmentación, donde en el paso de correspondencia de píxeles utilizan una medida de similitud que utiliza información sobre la intensidad de los píxeles, el gradiente en dirección horizontal y el gradiente en dirección vertical, y en la segmentación de la imagen en color utiliza la técnica de clasificación conocida como "*mean-shift*".

Alagöz (2008) propone un método basado en crecimiento de regiones. Se diferencia dos etapas, la primera etapa se llama proceso de selección de la raíz ("*Root Selection Process"*), en la que se busca un punto raíz a partir del cual se hace crecer la región; y la segunda etapa se llama proceso de crecimiento de la región (*"Region Growing Process"*), en la que la región crece de acuerdo a una reglas determinadas.

2.3.2. Técnicas basadas en las características

Frente a las técnicas basadas en el área en las que la correspondencia se realiza píxel al píxel, están las técnicas basadas en las características (Tang, Wu y Chen, 2002) en las que se emplean determinadas estructuras de un nivel superior al píxel para establecer correspondencias entre ellas. En concreto, se entiende por característica alguna estructura significativa de la imagen, en concreto: *puntos de borde*, *segmentos de borde* (rectos o curvilíneos) o *regiones*. Es práctica habitual asignar un vector de atributos $\boldsymbol{x}$, a cada característica, como se ha indicado en la sección 2.2.3.

Uno de los operadores más utilizados para extraer puntos de borde es la Laplaciana de la Gaussiana, y los atributos más frecuentes en este tipo de primitivas son el módulo y dirección del gradiente y en ocasiones la Laplaciana, cuyos valores se obtienen por aplicación de los correspondientes operadores de gradiente y Laplaciana (Pajares y Cruz, 2007).

Como alternativa a los puntos de borde aislados surge el uso de puntos de borde conectados: *segmentos de borde rectos y curvilíneos*. Los algoritmos basados en bordes para escenas industriales, suelen ser más robustos que los métodos basados en el área, ya que centran su atención en los lugares donde existe un elevado contenido de información (Lane, Thacker y Seed, 1994).

La aplicación de la restricción de *conectividad*, que supone que los puntos de borde conectados en una imagen deben corresponderse con puntos de borde también conectados en la otra imagen, permite que la correspondencia sea más efectiva. Este tipo de características ha sido estudiado en términos de fiabilidad (Ayache, 1991, Kim y Park 1994, Breuel 1996) y robustez (Wuescher y Boyer 1991). Para segmentos de borde rectos se suele utilizar una generalización de los mismos atributos definidos para puntos de borde, obteniendo como valor final el promedio de los valores para cada punto a lo largo de todos los puntos que constituyen el segmento de borde, si bien cuando los segmentos de borde son curvilíneos el atributo de dirección no es aplicable.

Como alternativa a los bordes y segmentos de borde aparece el uso de las regiones, de las que se suelen utilizar atributos tales como: valor medio del nivel de

gris, área, rectángulo mínimo que contiene a la región, centroide, longitud del perímetro, eje principal (EP), ancho a lo largo del EP en píxeles, altura perpendicular al EP en píxeles, razón ancho/alto, color, momentos invariantes de Hu o Tsirikolias-Mertzios (Pajares *et al*., 2007)

En cualquier caso, estos atributos pueden ser simples, tales como el color de los píxeles (Klaus *et al*. 2006) o propiedades obtenidas aplicando algún operador como el módulo del gradiente (Klaus *et al*. 2006) y el ángulo del gradiente o la Laplaciana (Lew *et al*. 1994). Ello a pesar de que en algunos contextos tanto el gradiente como la Laplaciana podían llegar a presentar cierta sensibilidad al ruido. En realidad, estos operadores tienen en cuenta los píxeles y sus vecinos; por tanto, desde este punto de vista podrían considerarse como basados en el área. El color es otro de los atributos que puede utilizarse a nivel de píxel individual o con la intervención de un entorno de vecindad (Klaus *et al*. 2006).

Los métodos basados en las características utilizan normalmente conjuntos de píxeles con atributos similares, ya sean píxeles pertenecientes a bordes (Tang, Wu y Chen, 2002; Grimson, 1985; Ruichek y Postaire, 1996), los bordes mismos (Medioni y Nevatia, 1985; Pajares y Cruz, 2006; Scaramuzza *et ál.*, 2008), regiones (McKinnon y Baltes, 2004; Marapane y Trivedi, 1989) o enfoques jerárquicos (Wei y Quan, 2004) donde primero se establece la correspondencia entre bordes o esquinas y después las regiones. En Tang, Wu y Chen (2002) se utilizan regiones con los tres siguientes atributos específicos para correspondencia: área, centroide y ángulos.

Existen numerosos trabajos en los que se utilizan los atributos anteriores para correspondencia mediante la aplicación de la restricción de semejanza. En Chehata *et ál.* (2003) dichos atributos se materializan en los siguientes: área, *bounding box*, momentos estadísticos espaciales. En Kaick y Mori (2006), aunque bajo un contexto de clasificación, se utilizan los momentos estadísticos de primer y segundo orden en el espacio de color *HSI*; estos atributos se obtienen a partir de los histogramas. En Renninger y Malik (2004) también se aplican descriptores de texturas, tales como los bancos jerárquicos de filtros. En Hu y Yang (2008) y Premaratne y Safaei (2008) se han aplicado satisfactoriamente momentos invariantes, donde se concluye sobre la

conveniencia de su uso en la correspondencia basada en regiones por la mejora en la exactitud a la hora de obtener las disparidades. En López y Pla (2000) se propone un método basado en grafos para tratar con errores de segmentación en correspondencia basada en regiones. Los nodos en el grafo son los posibles pares potenciales de correspondencias, mientras que a los arcos se les asigna valores teniendo en cuenta una medida de similitud entre las regiones bajo correspondencia.

En Wang y Zheng (2008) se extraen las regiones mediante un algoritmo de segmentación basado en el color y los píxeles pertenecientes a las regiones son emparejados obteniéndose un mapa de disparidad, que luego es refinado aplicando optimización cooperativa mediante el ajuste de algunos parámetros en las disparidades de las regiones segmentadas. En Ansari, Masmoudi y Bensrhair (2007) el color también se emplea para segmentar las regiones.

En Scaramuzza *et ál.* (2008) se utilizan líneas verticales como características en imágenes omnidireccionales y se calcula un descriptor invariante a rotaciones.

CAPÍTULO 3

3. Selección y descripción de métodos de correspondencia

3.1.Introducción

Tras explicar en el capítulo dos los conceptos necesarios que sustentan el presente trabajo, así como los principales métodos o técnicas encontradas en la literatura para determinar la estructura tridimensional de la escena, en este capítulo se explican las razones que nos han conducido a la elección de los métodos de correspondencia. Tras lo cual se explican con detalle el funcionamiento y comportamiento de los métodos seleccionados, que por otra parte son los candidatos para su aplicación en los métodos involucrados en los proyectos de investigación mencionados en el capítulo primero relativos al grupo ISCAR.

Lo primera argumentación estriba en argumentar el por qué la elección se ha orientado por las técnicas de correspondencia basadas en el área, y no se ha optado por ninguna de las basadas en las características. La explicación a ésta decisión está en la propia naturaleza de las imágenes objetivo del presente trabajo. Como se ha explicado en el capítulo uno, el presente trabajo tiene como meta la obtención de la escena tridimensional a partir de las imágenes tomadas por dos cámaras de ejes paralelos, con las que se captura el par estereoscópico, en entornos abiertos de exterior no estructurados, como puede ser el mar o la superficie de un planeta. En estos entornos no se dispone de características claras que puedan ser extraídas de las imágenes, a diferencia de entornos de interior como por ejemplo un pasillo o una habitación, donde se dispone claramente de puntos de borde, segmentos de borde o

regiones. Estas características en entornos cerrados como son los segmentos de borde provienen de la intersección de planos como las paredes y el suelo, techo u otras paredes; y las regiones se pueden obtener de puertas, marcos, mesas u otros objetos con texturas y colores semejantes. Todos estos elementos suelen ser muy comunes en entornos de interior y facilitan el uso de los métodos basados en características, ya que nos podemos beneficiar de toda esta abundante información abundante para aprovechar sus cualidades.

Al contener las imágenes, en las que se centra este estudio, pocos elementos característicos (puntos de borde, segmentos de borde o regiones) si se utilizasen métodos basados en características se alcanzaría un mapa de disparidad con muy escasa información, que al realizar una interpolación con el fin de obtener un mapa más denso nos proporcionaría datos de poca fiabilidad.

En consecuencia en base a lo anterior se ha decidido utilizar técnicas basadas en el área, y concretamente las siguientes: 1) la técnica basada en el coeficiente de correlación; 2) la técnica que hace uso de la energía global de error; 3) la técnica que utiliza segmentación y medida de similitud; y 4) la técnica que se fundamenta en el crecimiento de regiones. Esta última técnica podría ser considerada como basada en las características, si bien esta doble consideración será tratada cuando se comente la técnica más adelante.

Después de detallar cada uno de los algoritmos utilizados, en la sección 3.3 se explican los procedimientos destinados a mejorar el mapa de disparidad.

3.2.Métodos de similitud

3.2.1. Coeficiente de correlación

Esta técnica se basa en la utilización del coeficiente de correlación estadística, para determinar la correspondencia entre píxeles. Los parámetros estadísticos que se utilizan en esta técnica, se hallan sobre los valores de la muestra determinada por la ventana de vecindad, definida sobre el píxel a tratar. Al recorrer la imagen se abre una ventana centrada sobre el píxel de interés, constituyéndose un conjunto de valores (el valor del mismo píxel junto al valor de los píxeles que lo circundan) sobre los que se

hallan los parámetros estadísticos necesarios para el cálculo del coeficiente de correlación, como la media y desviación típica.

Para calcular la disparidad de un píxel de la imagen derecha, $P_d\ (x_d, y_d)$, se abre su ventana asociada, hallando la media y varianza de esta ventana. Siguiendo la línea epipolar determinada por dicho píxel se abren las correspondientes ventanas para los píxeles de la imagen izquierda, $P_i\ (x_i, y_i)$, que caen en esta línea y que por tanto son los hipotéticos candidatos, obteniendo también su media y varianza. Con estos valores se calcula el coeficiente de correlación para cada par de ventanas, siendo una ventana siempre fija, la de la imagen derecha, y la otra la que se forma tomando como elemento central el píxel de la imagen izquierda del par estereoscópico. De entre todos los valores del coeficiente de correlación obtenidos se elige el de mayor valor. De esta forma se determinan las componentes horizontales de los píxeles centrales de las dos ventanas que han generado ese valor mínimo y se obtiene el valor absoluto de la diferencia, $d = |x_i - x_d|$, que es directamente el valor de disparidad para dichos píxeles, representando realmente una medida de distancia entre las coordenadas de posición en el eje de *x* de las coordenadas de la imagen.

La expresión para hallar el coeficiente de correlación es la que se muestra a continuación:

$$C = \frac{\sigma_{ID}^2}{\sqrt{\sigma_I^2 \sigma_D^2}} \tag{3-1}$$

donde los subíndices *I* y *D* se refieren a las imágenes izquierda y derecha, respectivamente, σ_I^2 y σ_D^2 representan la varianza de los niveles de intensidad en las correspondientes ventanas y σ_{ID}^2 es la covarianza de los niveles de intensidad entre las ventanas izquierda y derecha. Estos coeficientes están definidos en las siguientes ecuaciones,

$$\sigma_k^2 = \sum_{i=1}^{M}\sum_{j=1}^{N}\frac{\left(I_k(i,j)-\mu_k\right)^2}{MN}; k = I,D \tag{3-2}$$

$$\sigma_{ID}^2 = \sum_{i=1}^{M}\sum_{j=1}^{N}\frac{\left(I_I(i,j)-\mu_I\right)\left(I_D(i,j)-\mu_D\right)}{MN} \tag{3-3}$$

en las anteriores expresiones la intensidad en cada píxel (*i,j*) viene dada por I_k (*i,j*) y la media del nivel de gris en el entorno de vecindad o ventana resulta ser μ_k . En ambos casos, el subíndice *k* indica si se trata de la imagen izquierda o derecha.

En la imagen izquierda se abren ventanas solamente sobre los píxeles que satisfacen las restricciones siguientes:

$$x_i > x_d$$

$$x_i < x_d + h$$

La primera restricción indica que para cada píxel de la imagen derecha, sólo se buscará correspondencia con píxeles de la imagen izquierda que se encuentren más a la derecha que éste, esto es así porque en la imagen izquierda los objetos de la escena están desplazados hacia la derecha respecto a su localización en la imagen derecha, se trata de una restricción derivada de la propia geometría del sistema estereoscópico. La segunda restricción establece una limitación a la hora de comprobar la correspondencia entre un píxel de la imagen derecha con un píxel de la imagen izquierda, estableciendo un límite superior para el valor posible de la disparidad. En este último caso, ya sea porque los objetos contenidos en la escena que se está analizando suelen situarse dentro de un rango aproximado de distancias, y por lo tanto también lo estará su disparidad, o simplemente por cuestiones de rendimiento, no interesa comprobar la correspondencia entre un píxel de la imagen derecha, y todos los que se encuentren a la derecha de éste sobre la línea epipolar de la imagen izquierda. El límite superior anteriormente mencionado se establece a priori, y por lo tanto se debe tener cierto conocimiento sobre la escena que se quiere analizar en

relación con la geometría del sistema antes de su procesamiento, de otro modo se corre el riesgo de perder información relevante. Si se establece el límite superior a un valor d_1 y se intenta hallar la disparidad de un objeto cuyo valor es mayor, d_2, la técnica actual le asociará al objeto el valor de la disparidad que tenga un mayor coeficiente de correlación, si bien como no se ha llegado a analizar la disparidad d_2, se le asignará un valor de disparidad menor y erróneo.

3.2.2. Minimización de la energía global de error

Este segundo método fue propuesto por Baykant (2008), en el cual se trataba la obtención de mapas de profundidad desde imágenes estéreo en color. La principal aportación de este método es la utilización de una matriz de *Energía de Error* para cada posible valor de disparidad. Esta matriz se construye mediante técnicas de correspondencia de bloques de píxeles. Se parte de la premisa de que las dos imágenes del par deben estar codificadas en el formato RGB, donde la imagen izquierda se representa como *L*(i,j,k) y la derecha como *R*(i,j,k), siendo ambas matrices de tres dimensiones. La matriz de *energía de error* se expresa también como una matriz tridimensional, más exactamente *e*(i,j,d), y se calcula mediante la expresión siguiente,

$$e(i,j,d)=\frac{1}{3\cdot n\cdot m}\cdot\sum_{x=i}^{i+n-1}\sum_{y=j}^{j+m+1}\sum_{k=1}^{3}\left(L(x,y+d,k)-R(x,y,k)\right) \qquad (3\text{-}4)$$

donde, k representa los diferentes canales de lo que está constituido el formato de imágenes RGB, pudiendo toma un valor del conjunto {*r, g, b*}, siendo *r* el valor para el canal rojo, *g* el valor para el canal verde y *b* el valor para el canal azul; i y j representan respectivamente el número de fila y columna del píxel perteneciente a la escena a tratar; y finalmente *n* y *m* son las dimensiones de la ventana que se utiliza para determinar la correspondencia de bloques. De esta forma *e*(i, j, d) expresa la energía asociada al píxel (i, j) suponiendo que su disparidad fuese *d*.

El criterio para elegir las correspondencias por este método consiste en escoger para cada píxel de la imagen de referencia el valor de disparidad que minimice su energía en relación a sus píxeles homólogos. Un sistema con una elevada energía es

altamente inestable y tiende a reducir la misma para llegar a un estado lo más estable posible; de esta forma por analogía con el mundo real el objetivo consiste en minimizar la energía de error total de la imagen, escogiendo para cada píxel el valor de disparidad que proporcione una menor energía para contribuir de este modo a reducir la energía global.

Para evitar cambios bruscos en la variación de energía entre píxeles contiguos de la imagen, posiblemente debidos a un error en el proceso de correspondencia, se aplica el filtro de la media aritmética sobre la matriz de las energías obtenida anteriormente, antes de escoger el valor de la disparidad para cada píxel. Este filtro, que suavizará la energía, viene dado por la siguiente ecuación,

$$\tilde{e}(i,j,d) = \frac{1}{n \cdot m} \cdot \sum_{x=i}^{i+n-1} \sum_{y=j}^{j+m-1} \left(e(x,y,d) \right) \tag{3-5}$$

Si el resultado no es satisfactorio, debido a que todavía existen variaciones muy pronunciadas en la matriz de energía, se puede volver a aplicar sucesivamente un determinado número de veces.

A modo de resumen y para visualizar de forma global el algoritmo, a continuación se exponen los principales pasos a seguir:

- Para cada valor de disparidad que se encuentre en el rango de búsqueda, se construye una matriz de energía de error dimensional, en la que cada píxel tiene una energía asociada. Las matrices anteriores se agrupan en una matriz tridimensional e (i, j, d). Figura 3-1.

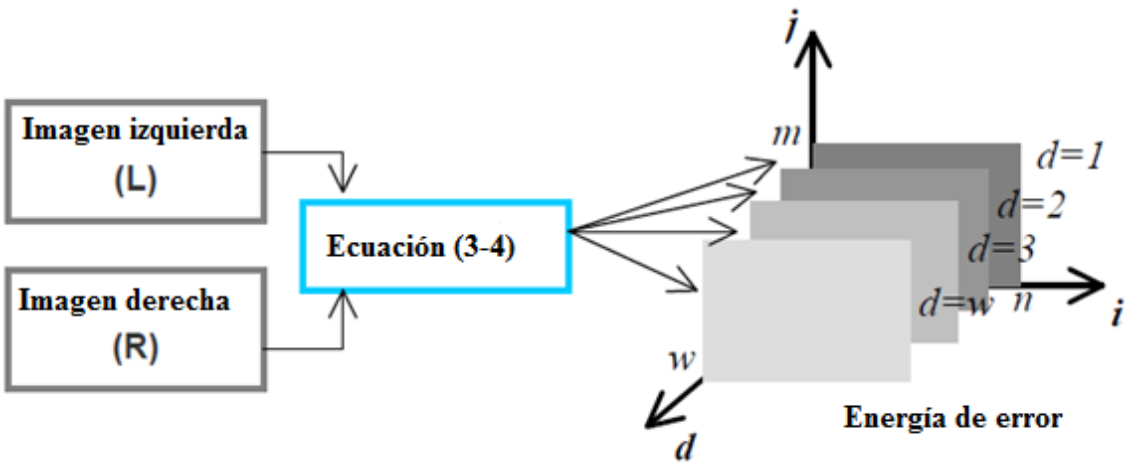


Figura 3-1: Construcción de las matrices de energía de error

- Realizar un suavizado mediante la aplicación del filtro de la media a cada una de las matrices de errores obtenidas en el paso anterior. Si no es suficiente con una pasada, porque existan grandes variaciones que no se corresponden con la realidad, se pueden realizar pasadas adicionales con filtro de la media. Obteniéndose finalmente la matriz $\tilde{e}(i,j,d)$. Figura 3-2.

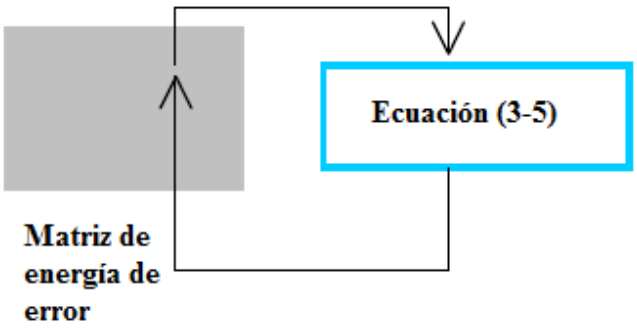


Figura 3-2: Suavizado de las matrices de energía de error

- Por último, para cada píxel con coordenadas (i, j) se localiza el valor de disparidad que minimiza la energía de error para ese píxel. Es decir, para un i y j dado se busca el valor d, que minimiza $\tilde{e}(i,j,d)$. El valor obtenido se asignará a dis (i, j), con lo que finalmente se dispondrá del mapa de disparidad del par estéreo originario en la matriz dis. Figura 3-3.

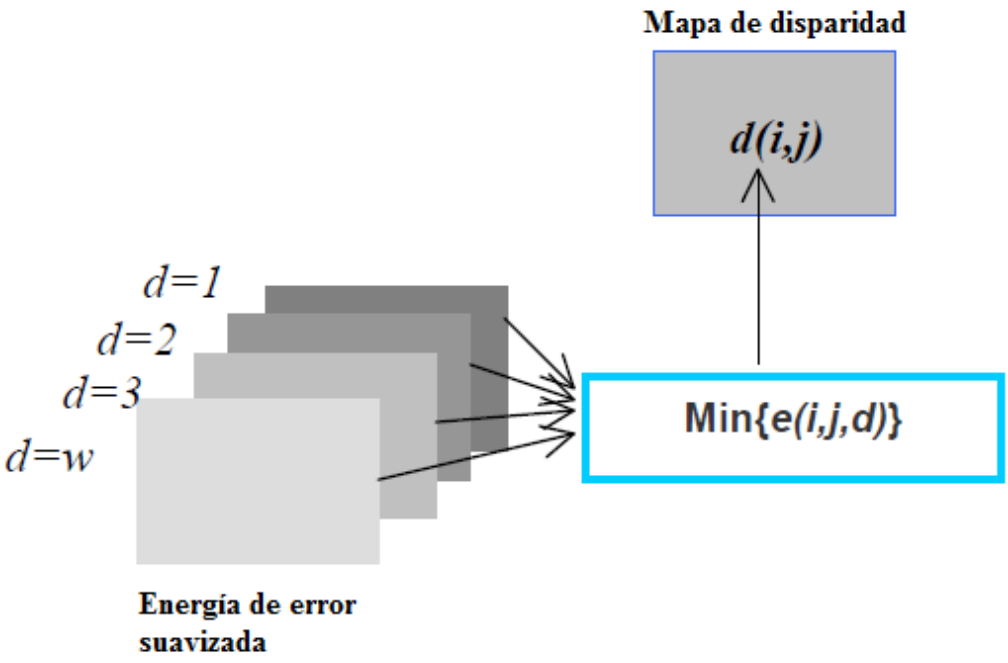


Figura 3-3: Construcción del mapa de disparidad por minimización de energía

3.2.3. Correspondencia basada en segmentación y medida de similitud

Este método está basado en la propuesta formulada por Klaus (2006), el cual se cimenta en la segmentación de la escena. Los fundamentos de la segmentación de la escena mantienen que, la estructura de la misma puede ser aproximada por un conjunto de planos no solapados en el espacio de disparidad y que cada uno de estos planos es coincidente con al menos una región de color homogénea de la imagen.

Los principales pasos en los que se fundamente el método son los tres siguientes:

1. se realiza una segmentación en color de la escena, localizándose regiones con un color homogéneo. Se asigna la misma etiqueta a cada uno de los píxeles que caen dentro de la misma región.
2. se realiza el proceso de correspondencia local entre píxeles, obteniéndose el mapa de disparidad base. Sobre este mapa inicial se efectúan operaciones para alcanzar un mapa de disparidad de mayor calidad.

3. se asignan disparidades a cada una de las regiones que fueron obtenidas en el primer paso.

El algoritmo que descompone la escena en regiones de color homogéneo, asume que lo valores de disparidad varían suavemente en el interior de estas regiones y que las discontinuidades únicamente se producen en los bordes de las regiones. Si se produjese una sobresegmentación de la imagen no surgiría mayor problema, ya que esto ayudaría a confirmar la suposición inicial en la práctica. Sin embargo, una escasa segmentación desembocaría en un mapa de disparidad final erróneo, ya que en el momento de asignar disparidades a las regiones se atribuirían a algunos píxeles disparidades incorrectas. El algoritmo que se utiliza para la segmentación es el mean-shift de Comaniciu (2002).

Para el proceso de correspondencia se necesita definir una ventana con la que asignar a cada píxel una determinada puntuación. Esta puntuación será una medida de similitud tal como el cuadrado o el valor absoluto de las diferencias de intensidades. La medida de similitud que utilizamos, en nuestra versión, es una combinación de diferencias de intensidades en valor absoluto y dos gradientes, uno a través de las filas (en la dirección del eje Y) y otro a través de las columnas (en la dirección del eje X). La combinación sigue la siguiente ecuación:

$$C(x,y,d) = \alpha \cdot C_{SID}(x,y,d) + \beta \cdot C_{GRAD}(x,y,d) \tag{3-6}$$

así se tiene que, $C_{SID}(x,y,d)$ es la componente para el píxel (x, y) que tiene en cuenta las diferencias de intensidad suponiendo una disparidad d; $C_{GRAD}(x,y,d)$ es la componente para el píxel (x, y) que tiene en cuenta los gradientes; los parámetros α y β tienen como finalidad establecer la contribución de cada coeficiente al cómputo final de la similitud. Si se desea que tenga un mayor peso la diferencia de intensidades se da un mayor valor a α; por el contrario si tiene una mayor repercusión el gradiente se da un mayor valor a β. Las expresiones descritas en mayor detalle para C_{SID} y C_{GRAD} están dadas por:

$$C_{SID}(x,y,d) = \sum_{(i,j)\in N(x,y)} \sum_{k=1}^{3} I_1(i,j,k) - I_2(i+d,j,k) \tag{3-7}$$

$$\begin{aligned} C_{GRAD}(x,y,d) = & \sum_{(i,j)\in N(x,y)} \sum_{k=1}^{3} \left| \nabla_x I_1(i,j) - \nabla_x I_2(i+d,j) \right| \\ & + \sum_{(i,j)\in N(x,y)} \sum_{k=1}^{3} \left| \nabla_y I_1(i,j) - \nabla_y I_2(i+d,j) \right| \end{aligned} \tag{3-8}$$

donde I_1 e I_2 se refieren a las imágenes que componen el par estereoscópico, $I_1(i,j,k)$ representa el valor que toma el píxel de coordenadas (i,j) en el canal k ; $N(x,y)$es la ventana definida para el píxel (x,y), todos los píxeles que caigan dentro de esta ventana serán tenidos en cuenta en las operaciones efectuadas para dicho píxel; ∇_x es el operador gradiente en la dirección del eje X, tomado de izquierda a derecha; ∇_y es el operador gradiente en la dirección del eje Y, tomado de arriba a abajo.

En el paso de correspondencia se hallan dos matrices, $C_1 y\ C_2$, de similitud, una tomando como imagen de referencia la derecha y otra tomando como imagen de referencia la izquierda, mediante la ecuación (3-6). A partir de estas dos matrices de todos los posibles valores se eligen los mejores, tras lo cual se obtiene un mapa de disparidad para cada valor de disparidad. El criterio de selección de los mejores se refiere a que se escogen los valores de disparidad que superen un cierto umbral y por tanto ofrezcan ciertas garantías de éxito como futuras correspondencias. Se consiguen así unos mapas de disparidad con valores con cierta fiabilidad. Una vez obtenidas las dos matrices de similitud, se construye para cada matriz un mapa de disparidad, $D_1\ y\ D_2$, siendo $D_k(x,y)$ el valor de d que minimiza la expresión $C_k(x,y,d)$. Ahora se construye un único mapa de disparidad de acuerdo a la siguiente regla de decisión:

$$D_F(x,y)=\begin{cases} D_1(x,y), & C_1\left(x,y,D_1(x,y)\right)<C_2\left(x,y,D_2(x,y)\right) \\ D_2(x,y), & C_2\left(x,y,D_2(x,y)\right)<C_1\left(x,y,D_1(x,y)\right) \\ \dfrac{D_1(x,y)+D_2(x,y)}{2}, & C_1\left(x,y,D_1(x,y)\right)=C_2\left(x,y,D_2(x,y)\right) \end{cases} \quad (3\text{-}9)$$

tras lo cual se continua con,

$$D_F(x,y)=\begin{cases} \infty, & \left|D_1(x,y)-D_2(x,y)\right|>\varepsilon \\ D_F(x,y), & C_2\left(x,y,D_2(x,y)\right)<C_1\left(x,y,D_1(x,y)\right) \end{cases} \quad (3\text{-}10)$$

La ecuación (3-9) indica que para cada píxel se escoge el valor de disparidad más fiable de entre las matrices D_1 y D_2, o una combinación de ellos. Por fiable se entiende la asociación entre píxeles más similares en función de las diferencias obtenidas. Si el valor de similitud para la disparidad de la matriz D_1 es menor que el asociado al valor de disparidad de la matriz D_2, entonces se opta por la disparidad de la matriz D_1. Sin embargo, si es menor la similitud asociada a la matriz D_2 se opta por la disparidad de la matriz D_2. En el caso de que sean iguales los valores de similitud, se opta por la media de los valores de disparidad de las matrices D_1 y D_2.

La finalidad de la ecuación (3-10) es aportar una mayor fiabilidad, ya que si la diferencia en valor absoluto entre la disparidad de las matrices D_1 y D_2, para un mismo píxel, es significativa se decide asignarle un valor representativo. Este valor representativo indica que el algoritmo no ha funcionado correctamente para ese píxel y se desconoce su disparidad.

En este momento todos los píxeles ya poseen un valor de disparidad que se ha acaba de determinar. Pero como a través del proceso de segmentación de regiones indicado previamente se parte de la hipótesis de que dentro de una región con color homogéneo la disparidad se mantendrá dentro de un rango de valores pequeño, se asignará el mismo valor de disparidad a estas regiones, evitándose de este modo valores espurios. Estos valores espurios en el cálculo de la disparidad son debidos a una determinación incorrecta por parte del algoritmo y provocan errores,

posteriormente, en el momento de utilizar el mapa de disparidad. Algunos de los errores mencionados anteriormente pueden provenir de la detección de falsos obstáculos en la navegación de los robots que lleven incorporado el sistema de visión estereoscópico o por la oclusión de objetos. La detección de falsos objetos, implica la identificación de un supuesto objeto en la escena cuando realmente no existe nada, lo que impondría falsas restricciones en una hipotética planificación de la trayectoria, complicando su cálculo además de producir abundantes errores. En el otro extremo se sitúa la oclusión de objetos, donde no se tiene conocimiento de la existencia de un objeto en la escena cuando realmente existe, de nuevo esto complica enormemente el cálculo de posibles trayectorias además de introducir errores de consideración. De la forma en la que el algoritmo trabaja, los mencionados objetos (falsos u oclusiones) se marcan como zonas de alto riego, de forma que en la navegación se consideran como tal, evitando problemas derivados de su falsa identificación.

En el último paso, gracias a la hipótesis inicial y con la finalidad de solucionar el problema planteado en el párrafo previo, se asignará un valor de disparidad a cada una de las regiones obtenidas tras la clasificación de las mismas. Todos los píxeles situados dentro de una misma región compartirán el mismo valor de disparidad. Como representante de estos valores se tomará el estimador estadístico de la mediana, para cada una de las regiones. Para hallar la mediana de la distribución de disparidad, intervienen todos los píxeles pertenecientes a la región dada pero cuyo valor de disparidad sea distinto de ∞, ya que estos valores indican que el algoritmo no ha funcionado correctamente y no se conoce la disparidad para los mismos.

Como último aspecto a comentar pero no por ello menos importante, resulta el hecho de que aunque este método se pueda considerar basado en las características por utilizar segmentación, es decir hacer uso de la información que proporcionan las regiones, y el gradiente, también se puede considerar basado en el área. Podemos realizar esta doble consideración porque las regiones no se utilizan para obtener atributos que formarán parte de vectores con los que obtener la similitud. Y si nos centramos en el gradiente, esta operación al tener en cuenta directamente los píxeles y sus vecinos podría ser catalogada como basada en el área. Además, como se comentará en el capítulo cuatro, puede resultar conveniente eliminar el proceso de

segmentación de la escena y la posterior asignación de un único valor a cada región. Esto se debe a la más que posible pérdida de información.

3.2.4. Correspondencia basada en líneas de crecimiento

El presente método sigue la idea propuesta por Lankton (2008), donde se describe un algoritmo basado en el crecimiento de regiones. El mecanismo de crecimiento de una región se divide en dos fases: en la primera fase se localiza el punto raíz a partir del cual se hace crecer la región, esta fase se denomina proceso de selección de la raíz (*Root Selectioin Process*); y en la segunda fase se hace crecer la región vinculada al punto raíz de acuerdo a una regla predefinida, esta fase se denomina proceso de crecimiento de la región (*Region Growing Process*).

Una región está constituida por todos los puntos que se asocian a un punto raíz. Así, para hacer crecer una región basta con incrementar el número de puntos asociados a un punto raíz. La regla utilizada para asociar un nuevo punto a un punto raíz es que la energía de error de un punto sea menor que un valor umbral de energía de error predeterminado. A este límite se le denomina umbral de la línea de crecimiento, en inglés "*LineGrowingThreshold*". Que un punto este asociado a un punto raíz significa que este punto tiene el mismo valor de disparidad que el punto raíz al que está asociado. En consecuencia con lo que se acaba de comentar, toda región constituida por asociación de puntos tendrá un valor determinado de disparidad.

Los pasos que se siguen a la hora de aplicar el algoritmo son los siguientes:

- **Paso 1 (Root Selection Process)**: Se selecciona un punto, el cual no pertenezca a ninguna región de crecimiento y se determina su disparidad usando la ecuación de la función de energía (3-4). Si no se encuentra ningún valor de disparidad cuya energía sea menor que el valor umbral, se selecciona el siguiente punto y se repite el paso actual. En caso de que se encuentre un valor de disparidad con una energía lo suficientemente pequeño, se le establece como el punto raíz actual, se crea la región vinculada con el punto raíz y se salta al paso 2.
- **Paso 2 (Region Growing Process)**: Se calcula la energía de error de todos los puntos vecinos del punto raíz actual con la misma disparidad que el punto raíz. Para cada punto vecino, se comprueba si su energía de error es menor que el umbral de error. En caso de que sea menor se

asocia el punto vecino al punto raíz actual (creciendo de esta forma la región), y en caso contrario se marca como libre.

- **Paso 3**: Se debe repetir el paso 2 hasta que la región actual deje de crecer. Cuando la región no pueda crecer más se vuelve al paso 1, para buscar un nuevo punto raíz y repetir todo el proceso de nuevo. Cuando todos los puntos de la imagen hayan sido procesados el algoritmo termina ya que el proceso habrá finalizado.

Para reducir la complejidad del algoritmo se restringe la dirección de crecimiento de las regiones, permitiendo que estas sólo puedan crecer en la dirección de las filas. Se puede imponer esta restricción sin miedo a que disminuya la calidad del mapa porque la disparidad de los píxeles de la imagen solo se produce en la dirección de las filas. Con esta simplificación los vecinos de un píxel se reducen a un único punto, inspeccionando solo el píxel siguiente al último píxel procesado para añadirlo a la región de crecimiento. De esta forma, las regiones de crecimiento se reducen a líneas de crecimiento.

3.3.Mejora del mapa de disparidad

Una vez que se tiene un mapa de disparidad construido con alguna de las técnicas anteriores o cualquier otra, es muy probable que éste contenga algunos errores. Estos errores pueden ser debidos a que se produjo una mala correspondencia en el proceso de correspondencia de características y no se asociaron correctamente las proyecciones de un elemento de la escena 3-D en sendas imágenes del par estereoscópico. Otra posible razón es que el mapa de disparidad final obtenido no sea válido debido a que alguna de las imágenes estaban modificadas, a pesar de que el algoritmo que obtiene la disparidad de la imagen se haya comportado correctamente. El origen de dicha modificación puede ser debido a que las imágenes originales o alguna de ellas contenían abundante ruido, introducido por cualquier factor durante su captura tal como vibraciones, o por la existencia de oclusiones.

Para solucionar estos problemas se realiza un proceso gracias al cual se intenta mejorar el mapa de disparidad. Este proceso de mejora puede ser de naturaleza local o global. En el proceso de naturaleza local se tiene en cuenta un pequeño número de píxeles vecinos (vecindad) para modificar el valor del píxel central de dicha vecindad.

En el proceso de naturaleza global se tiene en cuenta toda la imagen para modificar la disparidad de un píxel. La vecindad puede ser más o menos extensa, de forma que si el proceso es local y la vecindad grande el proceso puede llegar a ser global y viceversa.

La idea para llevar a cabo la mejora del mapa de disparidad estriba en el hecho de que las relaciones estructurales de la escena 3D se mantienen en la imagen. En efecto, una determinada estructura, tal como una textura, u objeto que contiene sus partes unidas en la escena, su imagen también preserva esta unión en la imagen 2D. Esto permite suponer que si un píxel está rodeado de píxeles con una determinada disparidad, él debería tener una disparidad similar a la que le rodea sobre la base de que todos ellos pertenecen a una misma estructura situada en una localización espacial dada y por tanto con valores de disparidad similares en la imagen.

Una vez argumentado el cambio de alcance en los métodos de naturaleza global se pasa a describir algunos métodos utilizados, tanto de naturaleza local como global.

3.3.1. Filtro de la media

Este método se basa en la aplicación de la media aritmética sobre una vecindad en el mapa de disparidad. La vecindad que comprenda la ventana es utilizada como espacio muestral para el cálculo de la media. La finalidad de este filtro es corregir el valor de la disparidad de un píxel con un valor muy alto o bajo respecto de todos los píxeles que lo circundan, ya que muy probablemente su valor sea debido a un error sobre la base de que todos ellos pertenecen a la misma zona con similar disparidad. Si todos los valores que rodean a este píxel tienen un valor de disparidad similar, la disparidad del píxel central se mantiene por la aplicación de la media. En la Figura 3-4 se muestra un ejemplo en el que se clarifica la aplicación del filtro de la media, utilizando una ventana de dimensión tres. El píxel de interés tiene un valor de disparidad de 10 y sus vecinos unos valores de 84, 85 y 86, esto hace suponer que probablemente el valor del píxel central sea erróneo. El filtro de la media calcula el promedio de los valores de los píxeles en la ventana y asigna este valor al píxel central. Utilizando una ventana de dimensión $n \times m$, se tendrán $n \cdot m - 1$ vecinos, que para el caso del ejemplo serán $3 \cdot 3 - 1 = 8$ vecinos. Como se puede observar en la Figura

3-4, tras aplicar el filtro, el píxel central pasa de tener un valor de disparidad de 10 a tener un valor de disparidad de 85.

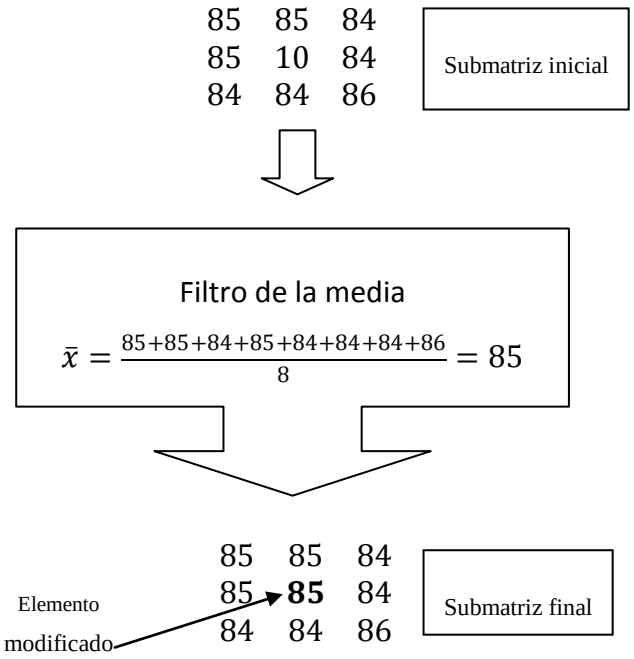


Figura 3-4: Aplicación del filtro de la mediana

3.3.2. Filtro de la mediana

Este método es muy parecido al filtro de la media, explicado en la sección anterior. La mediana es un valor de tendencia central, expresa el valor de la variable que ocupa la posición central de la distribución una vez que es ordenada de menor a mayor. Así para obtener la mediana de un conjunto de datos la primera operación a realizar es su ordenación de menor a mayor, y a continuación se procede a elegir el valor central de dicha ordenación. En caso de que el número de elementos sea impar se escoge el valor que ocupe la posición $\frac{n+1}{2}$, siendo n el número de elementos del conjunto ordenado, quedando la mediana como $M_e = x_{\frac{n+1}{2}}$. En caso de que el número de elementos sea par se toman los elementos que ocupan las posiciones $\frac{n}{2}$ y $\frac{n+1}{2}$, y se realiza la media entre los valores de dichas posiciones, siendo en este caso la mediana el valor $M_e = \frac{x_{\frac{n}{2}}+x_{\frac{n+1}{2}}}{2}$, que se asigna al píxel central de la ventana.

Este filtro se basa en la suposición de que puede existir más de un píxel erróneo en la ventana que define su vecindad, y que estos valores pueden ser mayores o menores que la media, pero que siempre el número de píxeles correctos será mayor que el de incorrectos. Por lo tanto, tras ordenar los píxeles de la ventana utilizada para modificar un píxel, excepto quizás unos pocos píxeles iniciales y finales la mayoría caerán en el centro de la distribución ordenada y consecuentemente se puede utilizar la media como un estimador para la corrección de la disparidad.

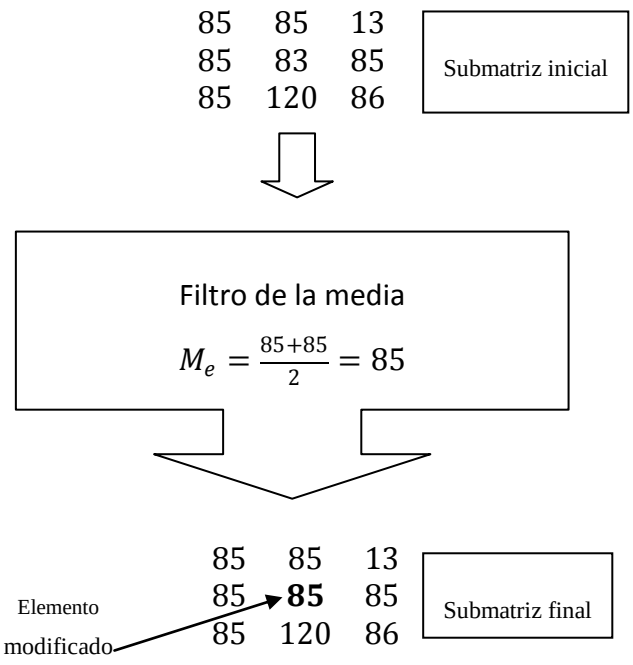


Figura 3-5: Aplicación del filtro de la mediana

A diferencia del filtro de la media, este filtro no se ve influido si existe algún píxel con un valor muy diferente al resto o incluso si hay varios. Pero para que este filtro se comporte correctamente debe haber suficientes elementos con valores acertados y semejantes. Al igual que en el filtro de la media se puede optar por considerar también el valor del píxel central y que se va a modificar.

3.3.3. Mapas Cognitivos Fuzzy

Los Mapas Cognitivos Fuzzy, conocidos en la literatura inglesa como Fuzzy Coginitive Maps (FCMs), son estructuras graficas borrosas para la representación del razonamiento causal. Fueron ideados por Kosko (1986), apareciendo por primera vez en 1986. Según el propio Kosko, la borrosidad de los FCMs permite distintos grados de incertidumbre sobre la causalidad entre objetos causales difusos. Una definición quizás más sencilla de entender sea la proporcionada por Kandasamy y Smarandache (2003): "Un FCM es un grafo dirigido formado por nodos y enlaces o aristas. Los nodos representan conceptos tales como políticas, eventos, valores, u otros y los enlaces representan causalidades, esto es relaciones causales entre conceptos, que vienen a determinar cómo un concepto o conceptos influyen en el resto. Los FCM se representan en forma de estructura de grafo, que permite representar una propagación causal sistemática, en particular encadenamiento hacia delante y hacia atrás, y permite que las bases de datos puedan crecer mediante la conexión de distintos FCMs.

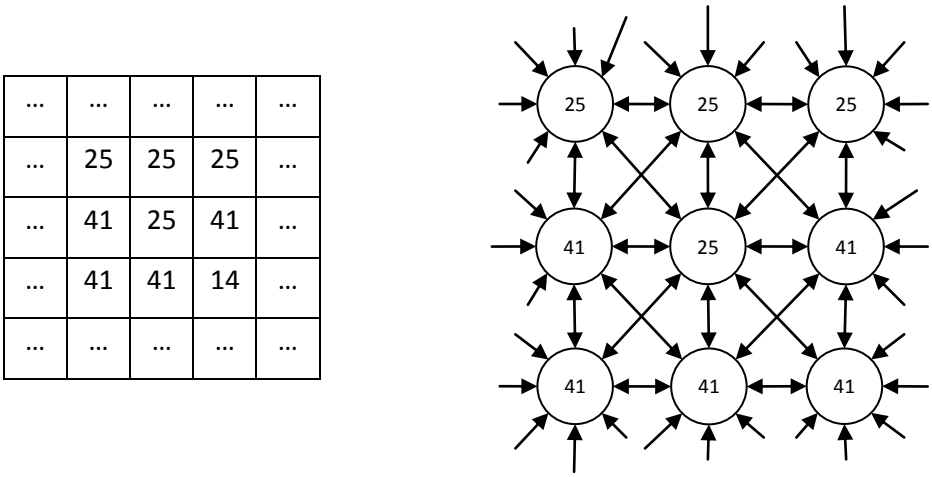
Kosko (1986) sostenía como idea que la mayoría del conocimiento es una especificación de clasificaciones y causas. En general las clases y las causas son inciertas (difusas o aleatorias), normalmente difusas. Esta falta de nitidez o borrosidad es trasladada a la representación del conocimiento y a las bases del conocimiento, donde se debe alcanzar un compromiso entre la adquisición y procesamiento del conocimiento. La borrosidad de la representación del conocimiento facilita la adquisición del conocimiento así como la concurrencia de las fuentes de conocimiento, pero dificulta el procesamiento (simbólico) del mismo. Con los FCMs se evita este compromiso.

El politólogo Robert Axelrod (1976) introdujo los mapas cognitivos para representar el conocimiento científico social. Estos mapas cognitivos son grafos dirigidos con signo. Donde los nodos son conceptos variables y las aristas son conexiones causales. El comportamiento de los mapas cognitivos se explica seguidamente. Una arista positiva desde el nodo A al nodo B significa que un incremento en el concepto que representa el nodo A tendrá como consecuencia un

incremento en el concepto que representa el nodo B, y una disminución del valor en el nodo A, conllevará una disminución del valor del nodo B. Una arista negativa desde el nodo A al nodo B, significa que un incremento en el valor del nodo A implica una disminución del valor en el nodo B, y una disminución de A implica un aumento del valor en B. En la mayoría de las ocasiones las relaciones de causalidad son difusas, indicándose distintos grados, y algunas veces con incertidumbre. De esta forma, con la introducción de la incertidumbre en los mapas cognitivos se llega a los mencionados Mapas Cognitivos Borrosos.

En definitiva y a modo de resumen los FCMs pueden ser vistos como redes neuronales usados para crear modelos de colecciones de conceptos unidos mediante relaciones causales que se establecen entre los conceptos (Tsardias y Margaritis, 1997; Tsardias y Margaritis, 1999; Kosko, 1986; Kosko, 1992; Miao y Liu, 2000). Estos conceptos son representados por nodos y las relaciones causales por arcos dirigidos entre nodos. Cada arco va acompañado de un peso causal que define el tipo de relación causal entre dos nodos.

Para el diseño de los FCMs se ha seguido el enfoque proporcionado por Herrera (2010) pero adaptado a nuestro caso, ya que en el citado trabajo se utilizaba un sistema estéreo diferente basado en lentes de ojo de pez para entornos forestales, donde la disparidad se mide en grados de ángulo. En el caso que nos ocupa la topología de la red es rectangular y de la misma dimensión que el mapa de disparidad. La red está constituida de tantos nodos como elementos tenga el mapa de disparidad, o lo que es lo mismo, por el número de píxeles por los que está constituida la imagen, ya que a cada píxel se le ha asignado un valor de disparidad inicial. Cada nodo de la red está localizado en la misma posición que el correspondiente píxel en la imagen derecha, y contiene como valor la disparidad asociada a ese píxel, que ha sido calculada previamente. De esta forma se obtiene una asociación entre píxeles de la imagen o imágenes y nodos de la red, pudiéndose aproximar un píxel a un nodo de la red, tal y como se representa en la Figura 3-6.



a) Mapa de disparidad

b) FCMs

Figura 3-6: FCMs asociado a un mapa de disparidad. a) Mapa de disparidad. b) FCM asociado al mapa de disparidad con todas sus conexiónes entre nodos

Según el esquema propuesto en este trabajo, sintetizado en la Figura 3-6, cada nodo de la red estará unido con todos los nodos pertenecientes a su vecindad, de suerte que según el tamaño de ésta la unión estará constituida por más o menos nodos. En el ejemplo de la Figura 3-6 se ha definido una vecindad de 3×3, por lo que cada píxel, a excepción de los situados en los bordes de la red, estará unido mediante enlaces con sus ocho vecinos más cercanos en las ocho direcciones posibles: arriba-abajo, izquierda-derecha y las dos diagonales principales. Los cuatro nodos situados en las esquinas de la red tendrán únicamente tres vecinos, y el resto de nodos que se encuentran en los bordes tenderán cinco vecinos. Para simplificar el dibujo se ha reduciendo el número de arcos que unen los nodos, se ha combinado en un solo arco de doble sentido los dos arcos, de un único sentido, que unen dos nodos.

Los nodos de la red neuronal así creada se caracterizan por tener asociado un valor de estado o nivel de activación. Estos estados se inicializan con los valores que proporciona el mapa de disparidad. Para cada nodo de la red se toma como valor inicial, el valor de disparidad que tenga el píxel del mapa de disparidad situado en el mismo emplazamiento que el nodo dentro de la red neuronal. Antes de de asignarle ese valor de estado se debe realizar una normalización de los valores de disparidad. Para poder utilizar los FCMs todos los valores de los nodos deben situarse en el rango [-1, 1], por tanto mediante un simple proceso de transformación lineal se puede pasar

del rango [0, d_{max}] al rango [-1, 1], donde d_{max} es el máximo valor de disparidad existente en el mapa de disparidad inicial.

Los valores que se asignan a los nodos tras la normalización de los valores de disparidad son lo que se llaman los estados o niveles de activación de la red neuronal. Una vez definida la topología de la red, inicializados los estados de los nodos y definidas las relaciones causales, se inicia un proceso iterativo de modificación de los estados de los nodos, que son reforzados o penalizados en las sucesivas iteraciones basándose en las influencias ejercidas por sus vecinos. No debe perderse de vista que la finalidad de utilizar los FCMs consiste en el suavizado del mapa de disparidad, y esto se consigue haciendo que la red evolucione hacia una configuración de estados en sus nodos lo más estable posible. Falta por definir los pesos causales w_{ik} que establecen las relaciones de causalidad entre los nodos. Estos pesos son los valores etiquetados a los arcos entre conceptos y toman valores en el intervalo fuzzy [-1, 1]. Gracias a estos pesos los FCMs son capaces de evolucionar, variando sus estados de activación debido a las influencias ejercidas entre los nodos. El peso causal w_{ik} expresa la influencia que ejerce el nodo i sobre el nodo k de la red, incrementando o decrementando el nivel de actividad del nodo afectado en función de su valor. Cuando el peso causal toma el valor cero, se dice que los nodos involucrados no se ejercen influencia mutua, es decir la relación de causalidad es nula. En los FCMs no existe realimentación de estados (Kosko 1986; Kosko 1992), lo que se traduce en el hecho de que no existe ningún arco con origen y final en el mismo nodo o dicho de otro modo podría considerarse un arco con causalidad nula, esto es $w_{ii} = 0$.

El objetivo final de los FCMs consiste en obtener un mapa de disparidad suavizado a partir de un mapa previo, que originalmente es el proporcionado por cualquiera de los métodos de correspondencia. La disparidad asignada a cada píxel es modificada en función de la influencia ejercida por sus nodos vecinos a través de sus estados de activación y la conexión que les une. En el estado inicial, iteración t=0, los valores de los estados de los nodos son los valores de disparidad normalizados, posteriormente tras sucesivas iteraciones *t* los estados de los nodos ven modificado su estado o nivel de activación de forma que se aproximen a los estados de los nodos vecinos con los que mantiene su influencia valores de disparidad. Existen diferentes

formas de modificación de los estados, siendo una de ellas la proporcionada por la siguiente ecuación, tomada de Tsardias y Margaritis (1997, 1999):

$$e_i(t+1) = f\left(e_i(t) + \sum_{k=1}^{q} w_{ki}(t) \cdot e_k(t)\right) - e_i(t) \cdot dec_i \qquad (3\text{-}11)$$

donde $e_i(t)$ es el estado de la neurona i en la iteración t en la red; $w_{ki}(t)$ es el peso causal asociado al arco que tiene su origen en el nodo k y su destino en el nodo i, en la iteración t; y por último dec_i es el llamado factor de decadencia para la neurona i, que toma valores en el rango [0, 1]. El factor de decadencia determina la fracción (en tanto por uno) del nivel actual de activación, que será sustraída del nuevo nivel de activación alcanzado. Este factor intenta reflejar la natural tendencia de la neurona a contrarrestar el efecto de cambio y acercarse el nivel de activación cero, así cuanto mayor sea el factor de decadencia mayor será el mecanismo. La función f se utiliza como función límite de forma que al actualizar el estado de un nodo éste no sobrepase el valor 1 o sea menor que el valor -1, que son los límites superior e inferior asignados a los estados de los nodos. En este trabajo se ha empleado la función hiperbólica que cumple con tales requisitos.

El valor *q* representa el número total de nodos existentes en la red, si bien debe tenerse en cuenta que los nodos que no pertenecen a la vecindad de i su peso asociado, $w_{ki}(t)$, es cero, por tanto los mismos no intervienen y no son tenidos en cuenta en el sumatorio, lo que reduce el coste computacional.

La ecuación anterior (3-11) representa el procedimiento de actualización del estado de un nodo, pudiendo expresarse de una forma más general como:

$$e_i(t+1) = f\left(e_i(t), A_i\right) \qquad (3\text{-}12)$$

donde $e_i(t)$ sigue expresando el estado del nodo i en la iteración t; A_i representa la influencia que ejercen todos los vecinos del nodo i sobre éste a través de sus arcos mediante los pesos causales; y por último *f* es una función que toma como argumentos el estado actual del nodo y las influencias de los vecinos (A_i).

Cada peso $w_{ki}(t)$ se define como un coeficiente de regularización que varía en cada iteración *t*. Teniendo en cuenta la asociación directa existente entre el píxel situado en la posición (x, y) y el nodo *i* en la red, la vecindad N_i^m contiene los *m* nodos que rodean al nodo *i*. Según esto se define el coeficiente de regularización en la iteración *t* como sigue:

$$w_{ki}(t) = \begin{cases} 1 - \left| e_i(t) - e_k(t) \right| si & k \in N_i^m \wedge i \neq k \\ 0 \qquad\qquad\qquad\quad si & k \notin N_i^m \wedge i = k \end{cases} \tag{3-13}$$

$$w_{ki}(t) = \begin{cases} -w_{ki}(t) & si\ signo\left(w_{ki}(t) \cdot e_k(t)\right) \neq signo\left(e_k(t)\right) \\ w_{ki}(t) & e.o.c. \end{cases} \tag{3-14}$$

El factor de decadencia se define basándose en la suposición de que una alta estabilidad en el estado de un nodo en la red implica que el nivel de activación para el nodo i podría desestabilizarse en una cantidad relativamente pequeña, llegando de nuevo a la estabilidad rápidamente, lo que de algún modo validaría dicho grado de estabilidad. Para aplicar este concepto se construye un acumulador de celdas con el mismo tamaño que la red, donde la celda *i* está asociada con el nodo situado en la misma posición y con idéntico nombre. Cada celda *i* contiene el número de veces h_i que el nodo *i* ha cambiado de forma significativa su nivel de activación. Inicialmente todos los valores h_i tendrán un valor de cero y a partir de entonces cada vez que se cumpla la relación $|e_i(t+1) - e_i(t)| > \varepsilon$ se incrementará en uno el valor h_i. La estabilidad de un nodo i se mide como la fracción de cambios acumulados en la celda *i* frente a los cambios en su vecindad N_i^m, y el número de iteraciones. De este modo el factor de decadencia se define de la siguiente manera,

$$dec_i(t) = \begin{cases} 0 \quad si\ h_i = 0 \wedge \overline{h_k = 0} \\ \dfrac{h_i}{\left(\overline{h_k} + h_i\right) \cdot t} \quad e.o.c. \end{cases} \tag{3-15}$$

donde $\bar{h}_k$ es la media de los valores acumulados por los nodos $k \in N_i^m$, h_i es el acumulador para el nodo i, y t es el número de iteraciones.

Como el factor de decadencia resta una pequeña fracción al anterior nivel de activación, el resultado de esta operación puede llegar a ser mayor o menor que 1 o -1 respectivamente. Como el nivel de activación debe restringirse al intervalo [-1,+1] cuando se produzca alguna de tales situaciones, el valor del nivel de activación se fija al límite de -1 o +1 según corresponda.

Por último sólo queda definir cuándo se considera que la red ha conseguido la estabilidad, en cuyo caso el proceso se considera finalizado. Idealmente se dice que se ha llegado a la convergencia cuando los estados de los nodos no varíen entre iteraciones sucesivas. Si bien, esta situación raramente se consigue por lo que el criterio de variación se establece de forma que se considera que un nodo es estable si entre dos iteraciones consecutivas la diferencia entre sus estados no sobrepasa un determinado valor de umbral, que puede fijarse a un valor relativamente pequeño, por ejemplo del orden de 10^{-4}, este concepto se aplica a todos los nodos. Si bien, en la mayoría de las ocasiones rara vez se consigue la convergencia por este procedimiento, siendo habitual detener el proceso de actualización cuando se haya alcanzado un cierto número de iteraciones.

CAPÍTULO 4

4. Análisis de resultados

4.1. Objetivo del análisis y descripción de las imágenes

Una vez seleccionadas las técnicas que se han creído más apropiadas para llevar a cabo el proceso de correspondencia estereoscópica con ellas se obtiene un mapa inicial de disparidad, procediendo posteriormente al procesado de dicho mapa con el fin de mejorarlo, en el presente capítulo se analizan los resultados ofrecidos tanto por los métodos que obtienen el mapa inicial como los que se utilizan para su mejora.

La implementación de los distintos métodos que se utilizan en este trabajo se ha realizado en Matlab (2010) y probados con la versión 2007b bajo un procesador AMD Turión X2 y Sistema Operativo Windows Vista. Se trata de un lenguaje científico interpretado. Debido a este hecho el coste computacional de los procesos implementados en Matlab es superior a otros lenguajes de distinta naturaleza, sin embargo a su favor está la facilidad en el manejo de matrices e imágenes así como el gran número de operaciones y funciones que ya posee. Debido a que el objetivo de este trabajo no consiste en estudiar los tiempos de cómputo sino el comportamiento de los métodos propuestos, la comodidad en el tratamiento de las imágenes ha sido determinante en su elección como lenguaje utilizado.

En primer lugar se analizan las técnicas comentadas en la sección 3.2, destinadas a obtener un mapa de disparidad sobre el que se aplicarán las técnicas de suavizado descritas en la sección 3.3. Para evaluar estas técnicas se hace uso del estimador conocido como Error Cuadrático Medio, con el que se comparará el mapa de disparidad obtenido con cada uno de los métodos y el mapa de disparidad

considerado como correcto, técnicamente conocido como "*ground-truth*" cuando éste esté disponible.

La finalidad de la elección de este tipo de análisis consiste en determinar el comportamiento de los métodos estudiados.

Tras evaluar las técnicas de la sección 3.2, se continua con la evaluación de las técnicas del apartado 3.3, de igual manera a como se ha hecho para las técnicas de similitud.

4.2.Descripción del conjunto de imágenes utilizadas

La imágenes que se han utilizado para valorar los algoritmos propuestos han sido obtenidas de la base de datos de Middlebury (2010) (Hirschmüller y Scharstein, 2007), escogiendo imágenes del conjunto del año 2005 y 2006. Cada par de imágenes estereoscópicas disponibles en dicha base de datos se presenta con distintas resoluciones espaciales, esto es existen distintos tamaños para el estudio realizado en este trabajo se escogieron las de menor tamaño, las que denominan "*ThirdSize*", con la única finalidad de determinar por un lado los tiempos de ejecución y por otro la validez de los algoritmos. Esto está motivado por la necesidad de desarrollar métodos para su futura implantación en sistemas de navegación autónoma principalmente. Como los algoritmos procesan las imágenes píxel a píxel cuanto menor sea el tamaño de la imagen menor será el número de operaciones necesarias para extraer la información, mientras que los resultados en cuanto a las disparidades se presumen suficientes para la navegación. Los tamaños de las imágenes que forman el par estereoscópico tienen todas ellas una altura fija de 370 píxeles y una achura que varía entre 413 y 465 píxeles, si bien dado un par ambas imágenes poseen el mismo tamaño. La distancia focal es de 3740 píxeles y la línea base 160 milímetros. Las imágenes están almacenadas en el formato PNG (Portable Network Graphics). Todas las imágenes han sido tomadas en interiores con la iluminación perfectamente controlada, lo que permite comprobar que las tonalidades de color aparecen con una clara nitidez. No presentan aberraciones ni distorsiones radiales debidas al sistema óptico utilizado, por lo que se supone que o bien has sido corregidas o el sistema óptico con el que estaban equipadas las cámaras en el momento de la captura era de alga calidad. Tampoco

poseen desplazamientos verticales, debidos al desalineamiento de las cámaras, por lo que la búsqueda de correspondencias se puede realizar considerando únicamente la componente horizontal.

Se decidió escoger este repositorio de imágenes, como núcleo de imágenes sobre las que realizar pruebas y analizar las técnicas por la sencilla razón de que cada par de imágenes posee su correspondiente mapa de disparidad, denominado "*ground-truth"* o "base de verdad". Este mapa de disparidad se supone que es el mapa de disparidad que se corresponde con la realidad, lo que es muy útil a la hora de evaluar la efectividad de los algoritmos estudiados. Gracias a la existencia de estos mapas de disparidad se puede cuantificar el error cometido por las técnicas empleadas, mediante algún estimador tal como el Error Cuadrático Medio (ECM).

La finalidad de la elección de este tipo de imágenes consiste en verificar el comportamiento de los algoritmos con el fin de aislar los problemas inherentes a los mismos, de tal manera que en el siguiente paso consistente en su aplicación al mundo real, sólo queden aquellos problemas derivados de esta naturaleza, tales como: distintos niveles de intensidad en las dos imágenes como consecuencia de brillos y otros efectos derivados de la iluminación en entornos no estructurados, distorsiones radiales de las lentes o desplazamientos verticales de las cámaras

Por cada par estéreo existen dos mapas de disparidad, uno que se realizó tomando como referencia la imagen izquierda, y el otro en el que se tomó como referencia la imagen derecha. Los mapas de disparidad también están en el formato PNG. En ellos el valor de intensidad asociado con un píxel se corresponde con el valor de disparidad del mismo, excepto cuando aparece el valor cero que indica un desconocimiento de la disparidad, bien porque el algoritmo de obtención de la misma falló en el momento de su obtención o bien por tratarse de zonas con oclusiones. Una consideración a tener en cuenta es que los mapas de disparidad de ThirdSize están multiplicados por un factor de tres, por lo que es necesario dividir el valor de cada píxel por esta magnitud para obtener un mapa de disparidad acorde con el tamaño de las imágenes.

Del conjunto total de imágenes se ha escogido un subconjunto constituido por diez pares estereoscópicos. Las imágenes que forman esta selección tienen por título: Aloe, Art, Books, Bowling1, Cloth2, Dolls, Laundry, Moebius, Wood1 y Wood2. El criterio seguido para realizar esta selección es el hecho de que las mismas presentan objetos muy próximos a las cámaras y por tanto con valores altos de disparidad. Esto es útil de cara a la navegación de los robots autónomos equipados con un sistema de visión estereoscópica ya que representan objetos cercanos que deben tratarse prioritariamente frente a los más lejanos con el fin de evitarlos. El valor máximo de disparidad que se emplea como límite de búsqueda de correspondencias entre las imágenes es el que viene dado por el máximo valor de disparidad entre todas las imágenes, resultando ser de setenta y siete píxeles. De este modo, los píxeles con altos niveles de disparidad podrían alcanzar su valor de disparidad real, si el algoritmo funciona correctamente, mientras que para los píxeles con bajos niveles de disparidad, los algoritmos intentarán hallar su disparidad dentro del rango más pequeño posible.

4.3. Análisis de resultados

Siguiendo con el planteamiento realizado en el capítulo tres, a continuación en la sección 4.3.1, se realiza la valoración de los resultados separando los obtenidos por los métodos utilizados en la sección 3.2, para posteriormente valorar los resultados en la sección 4.3.2 de los métodos de filtrado descritos en la sección 3.3.

4.3.1. Resultados de los métodos individuales

Como se ha comentado previamente se utiliza el criterio ECM para evaluar la calidad de los métodos, ya que disponemos de los mapas *ground-truth*. Si bien surge un cierto problema a la hora de calcular el ECM, se trata de la presencia de píxeles con valores de disparidad desconocidos. Tanto en los mapas que proporciona Middlebury (2010) como los generados por los cuatro métodos seleccionados, existen píxeles con disparidad cero, ya que esto supondría que el objeto o la zona del espacio del que procede el píxel se encuentra en el infinito, por ello el valor cero sirve para expresar el desconocimiento de la disparidad de un píxel. Teniendo en cuenta este hecho se calculará el ECM se calcula de dos formas distintas. En una de ellas no se tienen en

cuenta los píxeles con disparidad cero para calcular el ECM, ya sea en el mapa *ground-truth* o en el mapa generado por el método utilizado, y en la otra sí.

Como se ha mencionado previamente la existencia de un cero en el mapa de disparidad puede ser debida a una oclusión o a que el método no ha sido capaz de encontrar un valor de disparidad adecuado. Si se trata de una oclusión y el píxel contribuye en el cómputo del ECM se trataría de un error porque en realidad se desconoce la distancia al objeto y por tanto su disparidad. Por el contrario, si se trata de que el algoritmo ha fallado a la hora de determinar un valor de disparidad para ese píxel, no tendría que tener la misma consideración que una determinación errónea de la disparidad, ya que la zona se marcaría como peligrosa, evitando que el vehículo autónomo transite por ella. Un problema derivado de esta consideración es que se corre el riesgo de que los métodos que generen muchas indeterminaciones en el cálculo de la disparidad obtengan un valor pequeño del ECM, mientras un método que genera un número pequeño de indeterminaciones obtenga un valor alto del ECM, aún cuando se comporte mejor en términos del cálculo de la disparidad global. Como solución para evitar esta problemática, se decide que todos los píxeles contribuyan al ECM, es decir, aunque un píxel tenga un valor de disparidad de cero, se considera como un valor más, teniéndose en cuenta en el cómputo del ECM, esto supone el segundo procedimiento de cálculo del ECM.

Para cada par de imágenes estereoscópicas del conjunto de evaluación se ejecutarán los cuatro métodos de correspondencia explicados en la sección 3.2, más una versión del método que utiliza la segmentación y medida de similitud. Esta versión consistirá en eliminar la segmentación y el posterior filtrado de la mediana que se aplica a cada región. Los cinco métodos utilizados son: coeficiente de correlación, minimización de la energía de error, correspondencia basada en segmentación y medida de similitud (similitud con segmentación), correspondencia basada en similitud (similitud sin segmentación) y correspondencia basada en líneas de crecimiento.

La Tabla 4-1 muestra los resultados relativos al cómputo del ECM teniendo en cuenta que los píxeles con disparidad cero no contribuyen al valor del error. En la columna de la izquierda aparecen los nombres de los pares de imágenes

estereoscópicas utilizadas. En la fila superior se incluyen los respectivos métodos de correspondencia empleados.

	Similitud sin segmentación	Similitud con segmentación	Correlación	Energía de error	Línea de crecimiento
Aloe	18.18	22.67	67.32	85.32	4.74
Art	18.00	57.37	89.42	90.32	31.69
Books	6.60	13.18	109.29	59.99	73.04
Bowling1	12.50	50.36	244.95	234.29	196.13
Cloth2	3.35	18.74	50.02	22.16	43.21
Dolls	3.30	4.44	64.97	20.74	9.08
Laundry	16.66	25.23	48.81	81.05	62.90
Moebius	5.61	5.66	48.02	27.48	40.95
Wood1	6.39	31.43	64.60	34.49	74.79
Wood2	4.34	11.95	144.90	173.06	233.62

Tabla 4-1: Tabla con los ECM sin tener en cuenta los píxeles con disparidad cero

A la vista de los resultados mostrados en la Tabla 4-1, y de acuerdo con los valores del ECM obtenidos, se puede inferir las siguientes conclusiones:

- El método de similitud sin segmentación es el método que menor ECM produce en ocho pares de imágenes, mientras que en los otros dos restantes se sitúa en segundo lugar.
- El método de similitud con segmentación es el segundo método con menor ECM en siete pares imágenes, en dos de los pares se sitúa en tercera posición y en el par restante alcanza la primera posición.
- Las dos versiones del método de similitud mantienen un comportamiento uniforme, situándose en la primera o segunda posición.
- Los métodos de correlación, minimización de la energía de error y línea de crecimiento no mantienen un comportamiento uniforme en las pares de imágenes analizados, resultando cada vez en una posición distinta y generalmente a partir de la tercera posición.

Para sintetizar toda la información de la Tabla 4-1, los resultados del ECM obtenidos por cada método sobre los diez pares se promedian, obteniendo el valor medio para cada uno de los métodos, que se corresponden con los resultados mostrados en la Tabla 4-2.

Método	Similitud sin segmentación	Similitud con segmentación	Línea de crecimiento	Energía de error	Correlación
$\overline{ECM}$	9.49	24.10	77.01	82.89	93.23

Tabla 4-2: ECM promediado para cada método sobre el conjunto de las diez imágenes sin tener en cuenta los píxeles con disparidad cero

De acuerdo con los resultados mostrados en la Tabla 4-2 el mejor algoritmo es el de *similitud sin segmentación* con un promedio en el error cuadrático medio de 9.49, seguido por el de *similitud con segmentación* con una media de 24.10, y ya a una mayor distancia se encuentra el de *línea de crecimiento* con una media de 77.01. Por lo tanto se puede afirmar que los mejores métodos para hallar la correspondencia, atendiendo al criterio de no considerar los píxeles con disparidad cero, son los de similitud. Y de entre estos dos métodos, es preferible el que no utiliza segmentación por presentar un ECM menor. Por simplicidad, en la Figura 4-1 a continuación se muestran los datos de la Tabla 4-2 de forma gráfica.

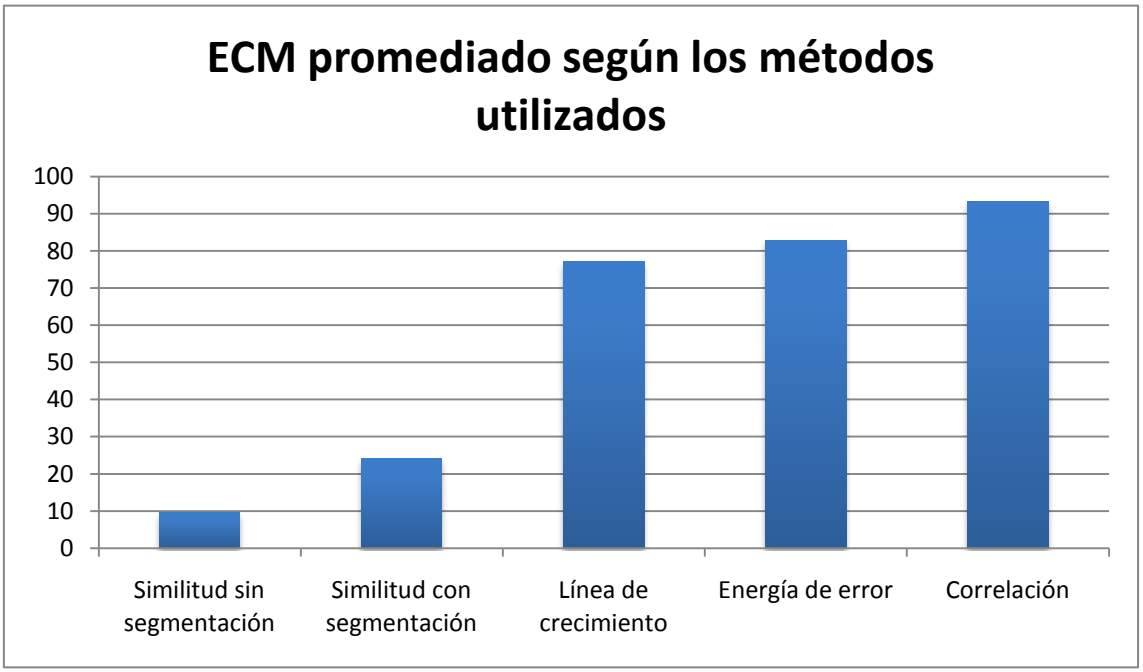


Figura 4-1: ECMs promediados, correspondientes al primer enfoque

A continuación evaluamos los resultados de los métodos mediante el segundo enfoque, esto es considerando los píxeles con disparidad cero para calcular el ECM. Como se ha mencionado previamente, el objetivo de incluir estos píxeles consiste

comprobar si influyen las indeterminaciones sobre la disparidad que producen algunos métodos o las oclusiones en el ECM.

	Similitud sin segmentación	Similitud con segmentación	Correlación	Energía de error	Línea de crecimiento
Aloe	306.12	59.42	103.95	123.20	394.61
Art	908.27	65.43	96.40	97.93	639.69
Books	1098.36	25.66	117.82	71.11	696.74
Bowling1	1203.97	115.99	267.02	262.46	455.73
Cloth2	687.66	44.84	69.03	45.36	1222.10
Dolls	730.75	13.20	76.85	33.16	794.57
Laundry	945.31	30.93	50.07	82.68	750.12
Moebius	538.44	15.42	53.16	33.71	680.63
Wood1	718.04	54.95	77.93	51.87	235.58
Wood2	966.57	17.68	153.29	179.57	337.60

Tabla 4-3: Tabla con los ECM teniendo en cuenta los píxeles con disparidad cero

A la vista de los resultados mostrados en la Tabla 4-3, se puede deducir lo siguiente:

- El método de *similitud con segmentación* es el que menor ECM produce en todas la imágenes menos en una, en la que es segundo por muy poco (tres unidades).
- Los métodos de *correlación* y de *energía de error global* comparten las posiciones segunda y tercera, situándose en ocasiones el método de correlación en la segunda posición y el de energía de error en la tercera y viceversa.
- Los métodos de *línea de crecimiento* y *medida de similitud sin segmentación* comparten las posiciones cuarta y quinta, situándose en ocasiones el método de la línea de crecimiento en la cuarta posición y el de medida de similitud en la quinta y viceversa.

Como en el caso anterior, con el fin de sintetizar la información dada en la Tabla 4-3, los resultados del ECM se promedian sobre el conjunto de los diez pares de imágenes, mostrando los correspondientes valores medios en la Tabla 4-4.

Método	Similitud con segmentación	Energía de error	Correlación	Línea de crecimiento	Similitud sin segmentación
$\overline{ECM}$	44.35	98.11	106.55	620.74	810.35

Tabla 4-4: ECM promediado para cada método sobre el conjunto de las diez imágenes teniendo en cuenta los píxeles con disparidad cero

A partir de los resultados mostrados en la Tabla 4-4 se deduce que el método de *similitud* aplicando segmentación es el que menor ECM produce en promedio, con un valor de 44.35; en segunda posición a una distancia superior al doble se encuentra el método *basado en la energía de error*, con un valor de 98.11; a continuación, se sitúa el método de *correlación* con un ECM de 106.55, siendo en este caso un valor muy próximo al de energía de error. Ya muy distanciados de éstos, se encuentra el método de *línea de crecimiento* con un ECM de 620.74 y el de *similitud con segmentación* con un ECM de 810.35. Se puede observar cómo el método de similitud resulta ser el mejor seguido a cierta distancia, entre el doble y el triple, por los métodos de energía de error y correlación. Estos dos métodos tienen un comportamiento muy similar y se diferencian en un ECM pequeño. Los métodos de línea de crecimiento y el basado en similitud sin segmentación son los que presentan un comportamiento errático mayor, en comparación con el método de correlación, el método que les precede, su ECM es aproximadamente seis y ocho veces superior respectivamente. A continuación, en la Figura 4-2 se muestran los datos de la Tabla 4-4 en forma gráfica para facilitar su visualización.

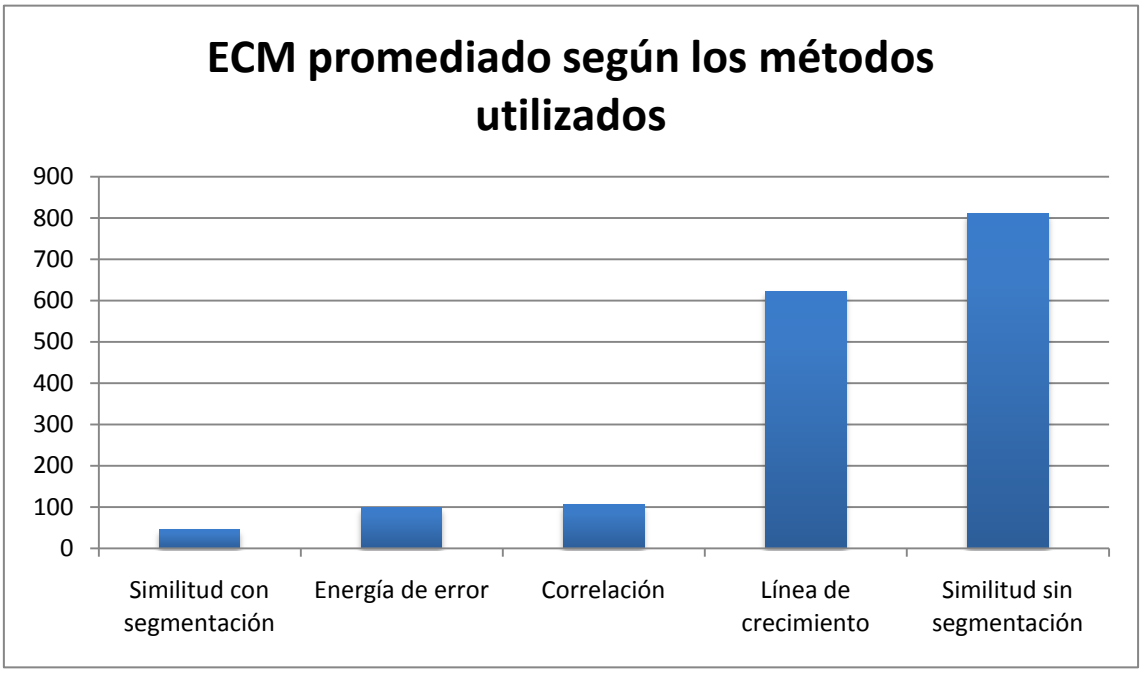


Figura 4-2: ECMs promediados, correspondientes al segundo enfoque

Como se ha expuesto previamente, existe una gran diferencia en los resultados obtenidos según el enfoque utilizado para calcular el ECM, por este motivo y tratando de aunar resultados, se ha considerado la conveniencia de utilizar otro criterio para su cálculo. Este nuevo criterio no tiene en cuenta los píxeles con disparidad cero en el ground-truth, pero si los píxeles con disparidad cero que presente el mapa de disparidad calculado por cada uno de los métodos. Aunque los resultados así obtenidos difieren de los anteriores, en líneas generales la valoración global de los métodos coincide con la obtenida por el segundo enfoque, de suerte que en este caso no existe aportación informativa adicional de relevancia, lo que nos permite quedarnos con la valoración de los anteriores.

Un análisis más detallado sobre los resultados proporcionados por los dos enfoques significativos, permite observar que los ECM mediante el primer enfoque (sin contabilizar los píxeles con disparidad cero) son menores que con el segundo enfoque (se contabilizan los píxeles con disparidad cero). Esto es ciertamente lógico porque aumenta el número de píxeles que participan en el cómputo del error. En cualquier caso, se puede comprobar que existen métodos que incrementan en mayor o menor medida el ECM. Así, los métodos de similitud con segmentación, correlación y energía de error global aumentan su ECM más suavemente, mientras que los métodos de línea de crecimiento y el de similitud sin segmentación presentan una diferencia más significativa, tal y como puede constatarse en los resultados mostrados en la Tabla 4-5, donde se muestran los porcentajes en los que se incrementa el ECM obtenido mediante el primer enfoque para cada uno de los cinco métodos.

Método	Similitud con segmentación	Correlación	Energía de error	Línea de crecimiento	Similitud sin segmentación
Incremento en % del ECM	84	14	18	706	8439

Tabla 4-5: Incremento en tanto por ciento del ECM

A partir de los resultados mostrados en la Tabla 4-5, los métodos con mayor crecimiento en el porcentaje de error son el de línea de crecimiento y el basado en la medida de similitud sin segmentación, en ambos casos sus ECMs mostrados en la Tabla

4-4 empeoran considerablemente con respecto a los resultados mostrados en la Tabla 4-2. Así, el método de línea de crecimiento pasa de estar situado en la tercera posición a la cuarta, que si bien sólo varía en una posición, mientras el valor de su ECM varía de forma significativa situándose en un 706%. El método de similitud sin segmentación desciende a la quinta posición desde la primera, pasando de ser el mejor al peor, mientras el valor de su ECM se incrementa en un 8439%.

El incremento del ECM tan brusco que experimenta el método de similitud sin segmentación se debe a que, sólo asigna un valor de disparidad cuando tiene un nivel de certeza adecuado en la decisión sobre esa disparidad. Por este motivo cuando el método de similitud determina un valor de disparidad este valor resulta ser suficientemente fiable, como consecuencia de ello el método se sitúa en la primera posición del ranking que utiliza el primer enfoque para calcular el ECM. Si bien, como deja muchos píxeles sobre los que no puede determinar su disparidad, al no alcanzar el nivel mínimo a partir del cual se está seguro en la asignación de un valor fiable, el ECM mediante el segundo enfoque alcanza un valor muy elevado, colocándose en la última posición de la clasificación. Algo parecido sucede con el método de la línea de crecimiento, aunque el mismo no ofrece tan buenos resultados como el método de similitud sin segmentación sin considerar los píxeles con valores de disparidad indeterminados.

Para respaldar la afirmación anterior se muestran en la Tabla 4-6 el número de indeterminaciones medio que genera cada método. En esta tabla se puede observar cómo los métodos de *línea de crecimiento* y *similitud sin segmentación* producen un elevado número de píxeles con valores de disparidad indeterminada.

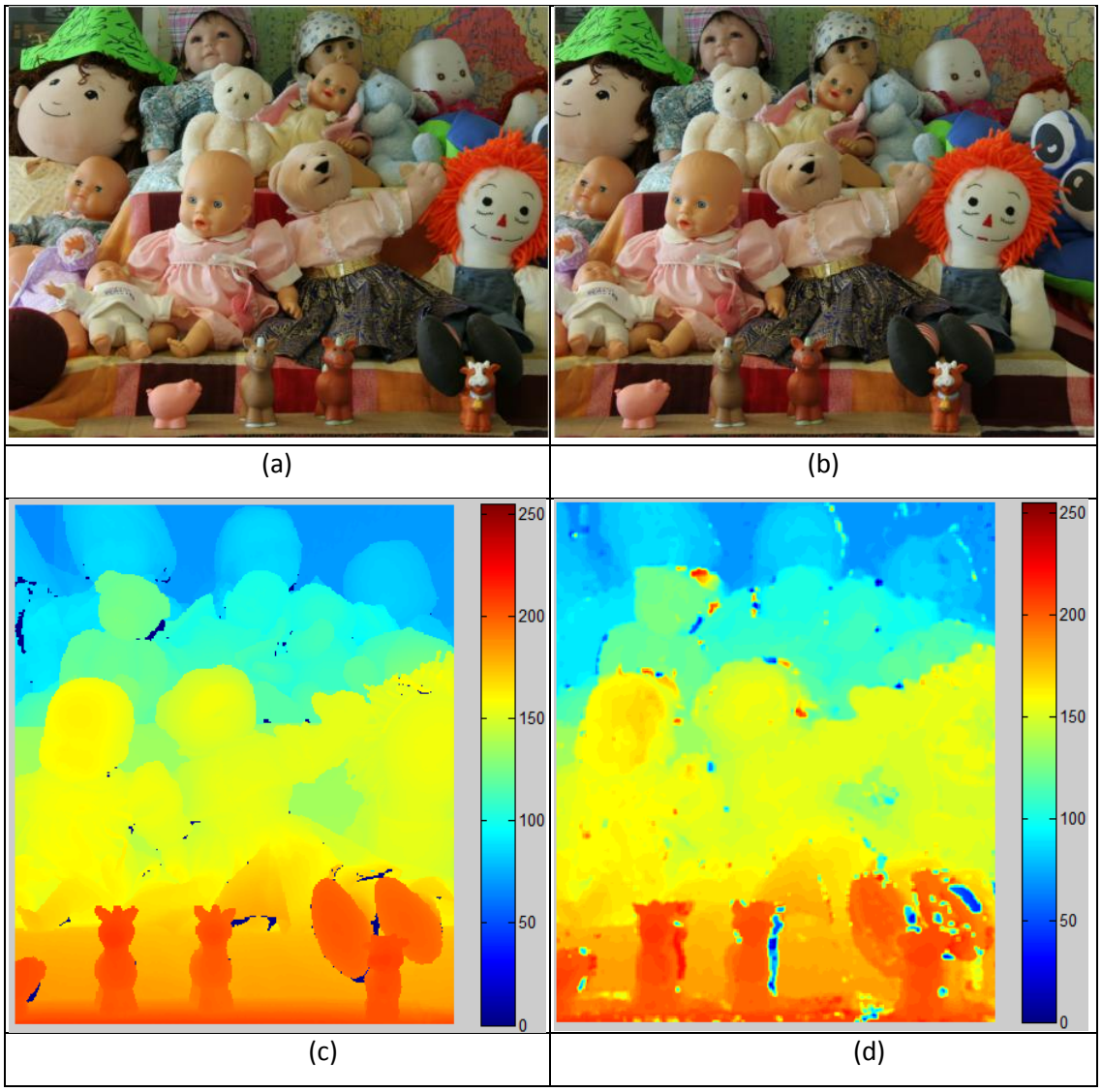
Método	Similitud sin segmentación	Similitud con segmentación	Correlación	Energía de error global	Línea de crecimiento
Píxeles con disparidad desconocida	42317.7	172.7	0	0	31348.8

Tabla 4-6: Número de píxeles medio con valores de disparidad indeterminada

A continuación, se realiza una valoración global sobre los métodos y filtros estudiados en la sección 3.3, a partir de los resultados obtenidos previamente.

A partir de los resultados obtenidos en las tablas anteriores se puede concluir que el mejor método es el de *similitud con segmentación*, ya que se sitúa en la segunda posición del primer enfoque para calcular el ECM y en la primera posición del segundo enfoque. En este caso no se ve la necesidad de aplicar filtrados sobre el mapa de disparidad. Esto es así, porque la finalidad del filtro es suavizar y uniformizar el mapa, y este método obtiene ya unos resultados suficientemente suavizados por el hecho de identificar regiones a las que asigna valores de disparidad similares a los píxeles que pertenecen a la misma región. Por tanto, se produciría un exceso de suavizado sin mejorar los resultados de forma significativa.

En la Figura 4-3 se muestran un ejemplo del mapa construido por el método de energía global de error, que sintetiza de forma global el comportamiento del método sobre el conjunto de imágenes analizadas. En (a) y (b) se muestran las imágenes estereoscópicas originales; en (c) el mapa de disparidad real ("*ground-truth*") de la escena obtenido junto con la imagen a partir de la base de datos de Middlebury (2010); y en (d) el mapa de disparidad elaborado por el método de energía de error global. La barra de color junto a los dos mapas de disparidad se corresponde con la representación de dichos valores en forma de color. Así cuanto más cálido (cercano al rojo) es el color, mayor es el valor de su disparidad, mientras que cuanto más frio (cercano al azul) sea el color, menor será el valor de su disparidad. Una simple inspección visual de los mapas de disparidad en relación al par estereoscópico original permite deducir sin mucha dificultad la evidencia de que a mayor disparidad más proximidad de los objetos con respecto al sistema estereoscópico que captura las imágenes. Entre ambos mapas de disparidad se observan ciertas diferencias que en términos generales no pueden considerarse significativas. Dichas diferencias son las que generan los porcentajes de error mostrados a lo largo de este capítulo. Estas mismas observaciones son extensibles al resto de imágenes analizadas.



4-3: Mapa de disparidad generado con el método de energía global de error

Según los resultados anteriores los métodos de *correlación* y el de *energía de error global* se perfilan como los más prometedores Los ECM de ambos son muy parecidos en ambos enfoques, siendo algo menor el correspondiente al de la energía, y además ninguno de los dos métodos genera píxeles con valores desconocidos significativos. Dada su similitud en cuanto a su conducta se podría seleccionar cualquiera de los dos, si bien hemos optado por el método de *energía global de error*, básicamente por su menor tiempo de ejecución, algo que aunque no se ha analizado con profundidad, se trata de un factor importante a tener en cuenta.

4.3.2. Resultados del filtrado sobre el mapa de disparidad

Una vez seleccionado el método de *energía de error global* se continúa con la aplicación de los distintos filtros descritos en la sección 3.3 sobre el mapa de

disparidad que genera este algoritmo. Para los filtros de la media y de la mediana se han utilizado ventanas de dimensión variable, concretamente con las siguientes dimensiones: 3×3, 5×5 y 7×7. Para el filtro basado en el FCM se ha considerado un único valor de vecindad en este caso de 3×3, ya que, como se explicó en la sección 3.3.3, tamaños de ventana mayores incrementan de forma excesiva el tiempo de cómputo además de producir un exceso de suavizado.

En la Tabla 4-7 se muestran los resultados del ECM promediado para el conjunto de las diez imágenes utilizadas en los experimentos y para los mapas de disparidad generados por cada filtro. Conviene recordar que el filtrado se ha realizado en todos los casos sobre el mapa obtenido mediante el método de energía global debido a los resultados que éste ha proporcionado. En esta ocasión los ECMs se han calculado considerando únicamente el segundo de los enfoques, que consistía en tener en cuenta todos los píxeles de la imagen, incluyendo los que tienen un valor de disparidad cero. En la columna de la izquierda aparecen los tamaños de las ventanas utilizadas y en la fila superior se incluyen los respectivos filtros empleados.

	Filtro de la media	Filtro de la mediana	FCM
Ventana de 3×3	91.56	93.56	93.06
Ventana de 5×5	84.85	86.88	-
Ventana de 7×7	80.17	80.58	-

Tabla 4-7: ECM promediado para cada filtro sobre el conjunto de los diez mapas de disparidad generados por el método de energía de error teniendo en cuenta los píxeles con disparidad cero

A partir de los resultados mostrados en la Tabla 4-7, se deduce que para una ventana de dimensión tres 3×3, el filtro con menor ECM es el de la media, seguido por el filtro que hace uso del FCM y por último el filtro de la mediana. Para las ventanas de dimensión cinco y siete, el filtro de la media sigue siendo el mejor, seguido por el filtro de la mediana, ya que el filtro del FCM no se evalúa con las vecindades de 5×5, ni de 7×7.

CAPÍTULO 5

5. Conclusiones

5.1.Conclusiones

En este trabajo se ha realizado una revisión de métodos de correspondencia estereoscópica con el fin de analizar sus posibilidades en términos de efectividad de cara a su futura incorporación en los sistemas reales descritos en la sección 1.2.2. El interés se ha centrado en el análisis de resultados de error.

Se ha utilizado un conjunto de imágenes procedentes de la base de datos de Middlebury (2010) con el fin de verificar su eficacia, al disponer de imágenes de prueba que permiten cuantificar los errores cometidos.

Tras el estudio realizado en el capítulo cuatro a partir de los resultados de correspondencia generados por los métodos descritos en las secciones 3.2 y 3.3, a continuación se exponen las conclusiones más relevantes derivadas de los mismos.

El mejor de los métodos resulta ser el basado en *segmentación y medida de similitud*. Además, es de los métodos con menor tiempo computacional tardando en promedio del orden de cuarenta segundos, sólo superado en velocidad por el método de *línea de crecimiento* que llega a situarse sobre los veintisiete segundos en promedio. Debido a que este método utiliza un proceso de segmentación por color, donde a todos los píxeles que pertenecen a una misma región se les asigna el valor de la mediana de la disparidad, no resulta beneficioso realizar un suavizado del mapa de disparidad. El principal problema que plantea este método es que los objetos existentes en la escena con una tonalidad uniforme y con distintos valores de disparidad son representados incorrectamente en el mapa final. A modo de ejemplo, si se intenta hallar la disparidad de una pared, con tonalidades de intensidad uniformes, que progresa a lo largo de un pasillo, el mapa de disparidad resultante muestra un

único valor de disparidad a largo de la pared, que se traduciría en colocar todos los puntos de la pared a la misma distancia. Evidentemente, esto plantea un problema importante no sólo desde el punto de vista de la navegabilidad en robótica sino desde el punto de vista de los errores que comete.

Podría pensarse en eliminar la parte de asignación de disparidades homogéneas, para posteriormente aplicar un suavizado mediante algunas de las técnicas de filtrado, si bien ahora el problema estriba en que en este caso deja un gran número de píxeles sin valores fiables de disparidad.

Ante los problemas anteriores, como siguiente opción podría pensarse en utilizar el método de *energía global de error* debido principalmente a su similar comportamiento con respecto al anterior, ya que simultáneamente no deja píxeles sin asignación de disparidad. Este puede ser el mejor candidato para continuar la investigación de futuro en entornos reales

En relación a los métodos de filtrado utilizados, se puede concluir que todos presentan un ECM parecido para una ventana de dimensión 3×3, si bien el tiempo de ejecución de los FCM supera al de la media y al de la mediana. Con ventanas de dimensiones 5×5 y 7×7 los resultados siguen siendo favorables al filtrado de la media, incluso respecto de la valoración de tiempos.

5.2. Trabajo futuro

La primera línea de investigación abierta es la derivada de los resultados obtenidos, sintetizados en la sección previa de conclusiones. En efecto, la investigación en los entornos reales debe continuar por la adaptación y modificación de los dos métodos más prometedores, esto es: *correspondencia basada en segmentación* y *energía de error global.*

Se planteó como objetivo el estudio de la efectividad de métodos, para lo cual se eligió una base de datos con posibilidad de prueba. Las imágenes disponibles en la base de datos utilizada, aunque reales, no plantean la problemática de las imágenes

reales en los entornos de exterior donde han de desarrollarse las aplicaciones derivadas de los proyectos mencionados en la sección 1.2.1, captadas con los sistemas estereoscópicos descritos en la sección 1.2.2. En base a los experimentos llevados a cabo hasta el momento sobre los mencionados sistemas estereoscópicos se han detectado los siguientes problemas que es necesario resolver, como paso previo o simultáneamente, relativos a la adaptación de los algoritmos de correspondencia identificados previamente.

- Las distorsiones radiales que se producen en las imágenes debidas a que los sistemas de lentes no son perfectos, ni se comportan de forma ideal, provocando aberraciones en las imágenes. En la imagen de la Figura 5-1 puede observarse este fenómeno, conocido también técnicamente como efecto barril. Se necesita un proceso de calibración de cámaras, si bien al tratarse de un tema ampliamente investigado en la literatura (Pajares y Cruz, 2007), el esfuerzo debe concentrarse en solucionar los problemas específicos relativos a los sistemas reales, tales como los mostrados en las Figura 1-1.
- Diferencias de intensidades en las imágenes capturadas por sendas cámaras. Esto proviene del hecho de que aunque las cámaras correspondan al mismo modelo y sean suministradas por el mismo fabricante, su comportamiento interno a veces difiere considerablemente, lo que provoca diferencias de intensidades significativas, que hacen que los algoritmos de correspondencia incrementen sus errores o incluso lleguen a fallar de forma llamativa.
- Desplazamiento vertical de las imágenes. Los ejes ópticos deben ser paralelos entre sí, y perpendiculares a la línea base, debe procurarse que el sistema esté perfectamente calibrado para cumplir con estos requisitos de la manera más fiel posible.
- Las cámaras deben alinearse verticalmente, de forma que entre las dos proyecciones de un mismo punto de la realidad tridimensional sólo exista un desplazamiento relativo horizontal o disparidad, a la vez que el desplazamiento horizontal es o nulo o mínimo.

A estos problemas se les suman otros debidos a la naturaleza propia de los entornos en los que los sistemas estereoscópicos deben trabajar, esto es, entornos de exterior. Las aplicaciones en exteriores trabajan con imágenes que presentan sombras y brillos debidos a múltiples fuentes de luz, elementos que producen reflexiones, iluminaciones no uniformes, etc. Estos brillos y sombras dificultan el proceso de correspondencia enormemente, ya que uno de estos elementos suele producirse en una imagen del par estereoscópico, estando ausente en la otra imagen del par. Los algoritmos de correspondencia deben adaptarse de forma que sean lo suficientemente robustos ante este tipo de circunstancias

(a) (b)

Figura 5-1: Par estereoscópico real

6. Bibliografía

Alagöz, B.B. (2008). Obtaining Depth Maps From Color Images By Region Based Stereo Matching Algorithms. OncuBilim Algorithm And Systems Labs. Vol. 08, Art.no. 04.

Ansar, A., Castano, A., Matthies, L. (2004). Enhanced real-time stereo using bilateral filtering, in: 2nd International Symposium on 3D Data Processing, Visualization, and Transmission, 455-462, Thessaloniki, Greece.

Ansari, M.E., Masmoudi, L., Bensrhair, A. (2007). A new regions matching for color stereo images, Pattern Recognition Letters, 28, 1679-1687.

Aschwanden, P., Guggenbuhl, W. (1993). Experimental Results from a Comparative Study on Correlation-Type Registration Algorithms. Robust Computer Vision. Forstner and Ruwiedel, eds., pp. 268-289, Wickmann.

Axelrod, R. (1976). Structure of Decision: the Congnitive Maps of Political Elites. Princeton, NJ: Princeton University Press.

Ayache, N. (1991). Artificial Vision for Mobile Robots: Stereo Vision and Multisensory Perception, MIT Press, Cambridge, MA.

Banks, J., Corke, P. (2001). Quantitative Evaluation of Matching Methods and Validity Measures for Stereo Vision. *Int'l J. Robotics Research*, vol. 20, no. 7.

Banno, A., Ikeuchi, K. (2009). Disparity Map Refinement and 3D Surface Smoothing via Directed Anisotropic Diffusion, IEEE 12th International Conference on Computer Vision Workshops, ICCV Workshops, pp. 1870-1877.

Barnard, S., Fishler, M. (1982). Computational Stereo, ACM Computing Surveys, 14, 553-572.

Barnea, D.I. y Silverman, H.F. (1972). A class of Algorithms for Fast Digital image Registration. IEEE Transactions Computers, 21, 179-186.

Baykant ALAGÖZ, B. (2008). Obtaining Depth Maps From Color Images By Region Based Stereo Matching Algorithms. OncuBilim Algorithm And Systems Labs. Vol.08, Art.No:04.

Bleyer, M., Gelautz, M. (2005). A layered stereo matching algorithm using image segmentation and global visibility constraints. ISPRS Journal of Photogrammetry and Remote Sensing, vol. 59, nr. 3, pp. 128-150.

Brown, M.Z., Burschka, D., Hager, G.D. (2003). Advances in Computational Stereo. IEEE Transactions on pattern analysis and machine intelligence, vol 25, no. 8, pp. 993-1008.

Breuel, T.H. (1996). Finding lines under bounded error, Patter Recognition, 29(1), 167-178.

Cassinelli, A., Perrin, S., Ishikawa, M. (2005). Smart Laser-Scanner for 3D Human-Machine Interface. CHI '05 extended abstracts on Human factors in computing systems, April 02-07, Portland, OR, USA.

Chan, T.F., Osher, S., Shen, J. (2001). The digital TV filter and nonlinear denoising. Image Processing, IEEE Transaction, vol. 10, pp. 231-241.

Chehata, N., Jung, F., Deseilligny, M.P., Stamon, G. (2003). A Region-Based Matching Approach for 3D-Roof Reconstruction from HR Satellite Stereo Pairs, in: Proc. VIIth Digital Image Computing: Techniques and Applications, Sun C., Talbot H., Ourselin S., Adriaansen T., Eds., Sydney, Australia, pp. 889-898.

Cochran, S.D., Medioni, G. (1992). 3-D Surface Description from binocular stereo, IEEE Transactions on Pattern Analysis and Machine Intelligence, 14(10), 981-994.

Comaniciu, D., Meer, P. (2002). Mean shift: A robust approach toward feature space analysis. IEEE: Trans. Pattern Analysis and Machine Intelligence, 24(5):603-619.

Deng, Y., Yang, Q., Lin, X., Tang, X. (2005). A symmetric patch-based correspondence model for occlusion handling. ICCV(International Conference on Computer Vision), pp. 1316–1322.

Grimson, W.E.L. (1985). Computational experiments with a feature-based stereo algorithm, IEEE Transactions on Pattern Analysis and Machine Intelligence, 7, 17-34.

Gehrig, S.K., Franke, U. (2007). Improving Stereo Sub-Pixel Accuracy for Long Range Stereo. Computer Vision, 2007. ICCV 2007. IEEE 11th International Conference on.

Herrera, P.J., (2010). Correspondencia estereoscópica en imágenes obtenidas con proyección omnidireccional para entornos forestales. Tesis doctoral. Facultad de informática. Universidad Complutense de Madrid.

Hirschmüller, H., Scharstein, D. (2007) Evaluation of cost functions for stereo matching. *IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR 2007),* Minneapolis, MN.

Hong, L., Chen, G. (2004). Segment-based stereo matching using graph cuts. IEEE Conference on Computer Vision and Pattern Recognition, pp. 508-515.

Hu, Q., Yang, Z. (2008). Stereo Matching Based on Local Invariant Region Identification, in: The International Symposium on Computer Science and Computational Technology, Volume 2, pp.690-693.

ISCAR (2010) (Ingeniería de Control, Sistemas, Automática y Robótica) http://www.dacya.ucm.es/area-isa/index.php?page=home)

Jin, H., Yezzi, A., Soatto, S. (2001). Variational Multiframe Stereo in the Presence of Specular Reflections. Technical Report TR01-0017, Univ. of California, Los Angeles.

Kaick, O.V., Mori, G. (2006). Automatic Classification of Outdoor Images by Region Matching, in: The 3rd Canadian Conference on Computer and Robot Vision (CRV'06), pp.9.

Kandasamy, V. and Smarandache, F. (2003). Fuzzy Congnitive Maps and Neutrosophic Cognitive Maps, ProQuest Information & Learning (University of Microfilm International.

Kim, D.H. and Park, R.H. (1994). Analysis of Quantization Error in Line-Based Stereo Matching, Pattern Recognition. 8, 913-924.

Klaus, A., Sorman, M., Karner, K. (2006). Segment-Based Stereo Matching Using Belief Propagation and a Self-Adapting Dissimilarity Measure. The 18th International Conference on Pattern Recognition (ICPR'06), vol. 3, pp. 15-18.

Kosko, B. (1986). Fuzzy cognitive maps. International Journal of Man-Machine Studies, Vol. 24, pp. 65-75, 1986.

Kosko, B. (1992). Neural Networks and Fuzzy System: a dynamical system approach to machine intelligence. Prentice-Hall, NJ.

Krotkov, E. (1989). Active Computer Vision by Cooperative Focus and Stereo, Springer-Verlag, Berlín.

Krotkov, E., Henriksen, K., Kories, R. (1990). Stereo ranging with verging cameras, IEEE Trans. Pattern Anal, Machine Intell., 12(12), 1200-1205.

Lane, R.A., Thacker, N.A., Seed, N.L. (1994). Stretch-correlation as a real-time alternative to feature-based stereo matching algorithms. Image and Vision Computing, 12(4), 203-212.

Lankton, S. (2010), http://www.shawnlankton.com/2007/12/3d-vision-with-stereo-disparity/ (disponible on-line).

Lew, M.S., Huang, T.S., Wong, K. (1994). Learning and feature selection in stereo matching, IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 16, pp. 869–881.

López, M.A., Pla, F. (2000). Dealing with Segmentation Errors in Region-based Stereo Matching, Pattern Recognition, 8(33), pp. 1325-1338.

Marapane, S.B., Trivedi, M.M. (1989). Region-based stereo analysis for robotic applications, IEEE Transactions on Systems, Man and Cybernetics 19(6), 1447-1464.

MATLAB. (2010). The Mathworks. http://www.mathworks.com/products/matlab/ (disponible online).

McKinnon, B., Baltes, J. (2004). Practical Region-Based Matching for Stereo Vision, in: 10th International Workshop on Combinatinal Image Analysis (IWCIA'04), Klette, R., Zunic, J., Eds., Springer, LNCS 3322, pp. 726–738.

Medioni, G., Nevatia, R. (1985). Segment Based Stereo Matching, Computer Vision, Graphics and Image Processing, 31, 2-18.

Miao, Y., Liu, Z.Q. (2000). On Causal Interference in Fuzzy Cognitive Maps. IEEE Transactions Fuzzy System 8(1) 107-119.

Middlebury (2010). http://vision.middlebury.edu/stereo/data/ (disponible on-line).

Moravec, H.P. (1977). Towards automatic visual obstable avoidance. Proceedings of the 5th International Joint Conference on Artificial Intelligence, MIT, Cambridge, Mass. Pp. 584-597.

Pajares, G., Cruz, J.M. (2006). Fuzzy Cognitive Maps for stereovision matching, Pattern Recognition, 39, 2101–2114.

Pajares, G., Cruz, J.M. (2007). Visión por Computador: Imágenes digitales y aplicaciones, 2ª ed., RA-MA, Madrid.

Pajares, G., Cruz, J.M., Aranda, J. (1998). Relaxation by Hopfield Network in Stereo Image Matching, Pattern Recognition, 31(5), 561 – 574.

Pajares, G., Cruz, J.M., López-Orozco, J.A. (1998). Relaxation labeling in stereo image matching. Patter Recognition, 33(2000), 53-68.

Pollard, S.B., Mayhew, J.E.W., Frisby, J.P. (1981). PMF: A stereo correspondence algorithm using a disparity gradient limit, Perception, 14, 449-470.

Premaratne, P., Safaei, F. (2008). Feature based Stereo correspondence using Moment Invariant, in: Proc. 4th Int. Conf. Information and Automation for Sustainability (ICIAFS'08), pp.104-108.

Reid, I.D., Beardsley, P.A. (1996). Self-alignment of a binocular robot, Image Vision computing, 14, 635-640.

Renninger, L.W., Malik, J. (2004). When is scene recognition, just texture recognition?, Vision Research, 44, 2301–2311.

Rohith, M.V., Somanath, G., Kambhamettu, C., Geiger, C. (2008). Towards estimation of dense disparities from stereo images containing large textureless regions, in: ICPR 2008, 19th International Conference on Pattern Recognition, 1-5, Tampa, FL.

Ruichek, Y., Postaire, J.G. (1996). A neural matching algorithm for 3-D reconstruction from stereo pairs of linear images, Pattern Recognition Letters, 17, 387-398.

Scaramuzza, D., Criblez, N., Martinelli, A., Siegwart, R. (2008). Robust Feature Extraction and Matching for Omnidirectional Images, Field and Service Robotics, Laugier, C., Siegwart, R., Eds., Springer, Berlin, Germany, Volume 42, pp. 71–81.

Scharstein, D., Szeliski, R. (2002). A taxonomy and evaluation of dense two-frame stereo correspondence algorithms, Int. J. Computer Vision, vol. 47(1-3), pp. 7-42.

Shirai, Y. (1987). Three-dimensional Computer Vision. Springer-Verlag, Berlín.

Solem, J.E., Aanæs, H., Heyden, A. (2007). Variational Surface Interpolation from Sparse Point and Normal Data. IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 29, no. 1.

Surveyor corporation (2010). http://www.surveyor.com/SRV_info.html

Tang, L., Wu, C., Chen, Z. (2002). Image dense matching based on region growth with adaptive window, Pattern Recognition Letters, vol. 23, pp. 1169–1178.

Tao, H., Sawhney, H.S., Kumar, R. (2001). A global matching framework for stereo computation. ICCV(International Conference on Computer Vision), pp. 532–539.

Tsardias, A.K., Margaritis, K.G. (1997). Cognitive Mapping and Certainty Neuron Fuzzy Cognitive Maps. Information Sciencies 101, 109-130.

Tsardias, A.K., Margaritis, K.G. (1999). An experimental study of the dynamics of certainty neuron fuzzy cognitive maps. Neurocomputing, 24, 95-116.

VIDERE Design (2010). http://www.videredesign.com/

Wang, D. (2005). The time dimension for scene analysis, IEEE Trans. Neural Networks, 16(6), 1401-1426.

Wang, Z.F., Zheng, Z.G. (2008). A region based stereo matching algorithm using cooperative optimization, in: IEEE Conference on Computer Vision and Pattern Recognition (CVPR'08), pp. 1-8.

Wei, Y., Quan, L. (2004). Region-Based Progressive Stereo Matching, in: Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'04), vol. 1, pp. 106-113.

Wuescher, D.M. y Boyer, K.L. (1991). Robust contour decomposition using a constraint curvature criterion, IEEE Transactions on Pattern Analysis and Machine Intelligence, 13(1), 41-51.

Zabih, R., Woodfill, J. (1994). Non-Parametric Local Transforms for Computing Visual Correspondence. Proceeding Third European Conference Computer Vision, pp. 150-158.